BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

#3

S&H Form: (2/01)

Attorney Docket No. 826.1730

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Hiroshi TSUDA

Application No.: Group Art Unit:

Filed: June 14, 2001 Examiner:

For: DOCUMENT COLLECTION APPARATUS AND METHOD FOR SPECIFIC USE, AND STORAGE MEDIUM STORING PROGRAM USED TO DIRECT COMPUTER TO COLLECT DOCUMENTS

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Assistant Commissioner for Patents
Washington, D.C. 2023l

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No. 2000-397966

Filed: December 27, 2000

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: June 14, 2001 By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001

PATANT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: December 27, 2000

Application Number: Patent Application
No. 2000-397966

Applicant(s): FUJITSU LIMITED

March 9, 2001

Commissioner,
Patent Office Kozo OIKAWA

Certificate No. 2001-3017566

日　本　国　特　許　庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２０００年１２月２７日

出願番号
Application Number: 特願２０００－３９７９６６

出願人
Applicant (s): 富士通株式会社

２００１年　３月　９日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号　出証特２００１－３０１７５６６

【書類名】　特許願
【整理番号】　0051888
【提出日】　平成12年12月27日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明の名称】　特定用途向けの文書収集装置、その方法及びコンピュータに実行させるためのプログラムを記録した記録媒体
【請求項の数】　10
【発明者】
【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
【氏名】　津田　宏
【特許出願人】
【識別番号】　000005223
【氏名又は名称】　富士通株式会社
【代理人】
【識別番号】　100074099
【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ
【弁理士】
【氏名又は名称】　大菅　義之
【電話番号】　03-3238-0031
【選任した代理人】
【識別番号】　100067987
【住所又は居所】　神奈川県横浜市鶴見区北寺尾７－２５－２８－５０３
【弁理士】
【氏名又は名称】　久木元　彰
【電話番号】　045-573-3683
【手数料の表示】
【予納台帳番号】　012542

【納付金額】　　21,000円

【提出物件の目録】

【物件名】　　明細書　1

【物件名】　　図面　1

【物件名】　　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　　要

【書類名】　　明細書

【発明の名称】　　特定用途向けの文書収集装置、その方法及びコンピュータに実行させるためのプログラムを記録した記録媒体

【特許請求の範囲】

【請求項１】　ネットワークから文書を収集する文書収集方法であって、

前記文書の参照関係に基づいて、前記ネットワーク上のコミュニティ内から文書を所定数以上収集し、

前記コミュニティ内から前記所定数以上の文書を収集した後、収集済み文書の参照関係に基づいて、前記コミュニティ内外から文書を収集する、

ことを特徴とする文書収集方法。

【請求項２】　前記収集済み文書の参照関係及びネットワーク上の場所を示す情報に基づいて重要度を算出し、

前記参照関係及び前記重要度に基づいて、収集すべき文書を決定する、

ことを特徴とする請求項１記載の文書収集方法。

【請求項３】　前記収集すべき文書は、前記コミュニティ内外別に決定される、

ことを特徴とする請求項２記載の文書収集方法。

【請求項４】　前記収集済み文書を検索した結果を、前記コミュニティ内外に分けて提示する、

ことを特徴とする請求項３記載の文書収集方法。

【請求項５】　ネットワークから文書を収集する文書収集方法であって、

ある分野に関する文書群である正例文書群と、前記分野と関連が少ない分野に関する文書群である負例文書群とを与え、

前記正例文書群及び前記負例文書群の参照関係に基づいて、前記分野に関する収集すべき文書を決定し、　前記ネットワークから前記収集すべき文書を収集する、

ことを特徴とする文書収集方法。

【請求項６】　前記参照関係に基づいて、前記正例文書群の文書からのみ参照される度合いである参照度を算出し、

前記参照度が高い文書を収集すべき文書として決定する、

ことを特徴とする請求項５記載の文書収集方法。

【請求項７】　前記参照関係に基づいて、前記正例文書群の文書を参照している収集済み文書から参照されている文書について、収集済み文書からの被参照数を示す共参照度を算出し、

前記共参照度が高い文書を収集すべき文書として決定する、

ことを特徴とする請求項５又は６記載の文書収集方法。

【請求項８】　前記負例文書群は、複数の分野に関する文書群の和集合である、

ことを特徴とする、請求項５乃至７記載の文書収集方法。

【請求項９】　前記収集済み文書で用いられている参照表現に基づいて、前記収集済み文書をまとめあげる、

ことを特徴とする請求項１乃至８記載の文書収集方法。

【請求項１０】　前記収集済み文書で用いられている参照表現に基づいて、前記収集済み文書にキーワードを付与する、

ことを特徴とする請求項１乃至９記載の文書収集方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書の収集に関し、特に特定用途に合わせて文書を効率的に収集する文書収集装置、その方法に関する。

【０００２】

【従来の技術】

イントラネット、WWW等のネットワーク上の文書の検索エンジンは、ネットワークから文書を収集する文書収集装置（ロボット）と、収集した文書用のキーワード索引を作成する検索エンジンとから実現される。

【０００３】

文書収集装置は、所与のネタＵＲＬ（Uniform Resorce Locator ）集（収集を開始する際の開始点となるＵＲＬ集）から文書収集を開始し、収集済みの文書か

らアンカー（参照関係）により参照されている未収集文書を次収集候補として収集し、といった処理を一定の回数繰り返すことにより動作する。このようにして文書収集ロボットは、数千万から数億のＵＲＬから文書を定期的に収集する。ここで、ＵＲＬとは、ネットワーク上の情報のありかと取得方法を指定する記述方式をいう。

【０００４】

ところで、今日、ネットワーク上の文書の増加スピードは速く、2000年1 月には、Inktomi 等によって、インターネットのユニーク文書は１０億文書に達したという調査結果が発表されている。また、2000年7 月には、アメリカCyveillance 社によって、インターネットの大きさは約２１億文書であり、2001年にはさらに倍の大きさになると予測されるという調査結果が発表されている。

【０００５】

１０億ＵＲＬから文書を収集するともなると、一日１００万ＵＲＬずつ（毎秒約１０ＵＲＬ＝４０Ｋバイト）収集したとしても収集し終わるには３年かかることになり、収集し終わった頃には最初の頃に収集した文書の情報は陳腐化していまう。そこで、用途に合わせて重要度の高い情報だけを効率よく収集する知的文書収集装置が求められていた。

【０００６】

特定用途の文書を優先して収集する文書収集装置には、以下のものがある。

・例えば、特開平9-311802に開示される発明のように、新しい情報を優先して収集する。

・内容が類似していると考えられる文書を収集する。その際に、以下の考え方を導入する。

【０００７】

ａ）階層数で収集範囲を制限する。

例えば、特開平9-218876に開示される発明のように、参照関係を有する文書は内容的にも近いと考えられるが、あまり階層的に離れると意味的な繋がりがなくなるため、階層数で収集範囲を制限して文書を収集するという考え方。

【０００８】

ｂ）意味的内容が近い文書のみ収集する。

例えば、特開平10-105572 に開示される発明のように、文書の中身のマッチングから意味的な近さを計算し、参照関係を有する文書のうち、意味的に近い文書だけを収集するという考え方。

【０００９】

ｃ）参照先を示す文字列が適当な文書のみ収集する。

例えば、特開平10-260979 及び特開2000-9011 に開示される発明のように、参照先を表している表現である参照表現、例えばＨＴＭＬであればアンカータグの内容に基づいて、その参照表現で参照されている参照先文書を次に収集するか否かを判定するという考え方。

・一般的に、より人気度の高い文書から優先して収集する。

【００１０】

被参照数の多い文書は、人気度が高いと考えられる。収集済みの文書群の文書から参照されている数（被参照数）が多い文書から順に収集することで、重要度の高い文書を優先して収集できるという考え方。

【００１１】

【発明が解決しようとする課題】

しかし、上述の従来技術の枠組みだけでは、企業のようなコミュニティのポータルサイトに求められるような文書の収集に用いるためには、不十分な点があった。例えば、企業内のポータルサイト、つまりコーポレートポータルの要件として、以下の点が要求される。

・社内外でリアルタイムに発生する膨大な文書を自動的に収集する。

・自動で意味解析及び分類分け（カテゴライズ）する。

・文書を収集し、分類した結果を画面の適当な場所に（人に合わせて）フィードする。

【００１２】

このうち、文書収集において、社内外の膨大な文書を漫然と収集するのではなく、文書の中から業務に関係するという観点から文書を選別して収集することが必要とされる。業務に関係するという観点は、特定の意味的内容を持つ、或いは

重要度を持つということとはやや異なる。例えば、ある程度の規模の企業が有するイントラネットコミュニティでは、文書内容も意味的に多様になるからである。また、社外（例えばインターネット）の文書は、趣味に関する情報も人気度が高くそうした情報は必ずしもコーポレートポータルにとって有用であるとは限らない。

【００１３】

しかし、従来の文書収集において用いられてきた枠組み、例えば、最新情報の優先取得、特定分野情報の優先取得、重要度優先取得という枠組みだけでは、このような趣味に関する情報のように、一般的に重要度が高いが必ずしもこのコミュニティにとって有用でない文書も収集されてしまうという問題があった。

【００１４】

また、例えば、上述の従来技術の「意味的内容が近い文書のみを収集する」と方法で文書を収集する場合、各々の考え方には以下の問題があった。

・単に階層数を予め制限する考え方は、処理は簡単であるが、本当に意味内容が近い文書を優先して収集しているのか、また、重要な文書を収集し逃していないのか、保証がない。

・文書の内容を比べて意味的内容が近いか否か判定する方式によれば、一般に自然言語処理を使って、文書に記載された本文を解析してキーワードを取り出し、取り出されたキーワードの類似度によって解析する。そのため、処理に時間がかかる。早くても、毎秒１００文書程度しか処理できない。従って、数十億ともいわれる文書を１つ１つ処理することは、現実的な時間内に行いがたい。また、そのように時間をかけて処理したとしても、その精度は７０から８０％程度である。さらに、この処理は、言語の種類に大きく依存するため、言語毎に判定ツールを備えることが必要となる。

・参照表現に基づいて収集するか否か判定する場合でも、参照表現で用いられる文字列には、「ホームページ」、「トップに戻る」及び「ここをクリック」といったような決まった語句（定番的ば語句）も多く、必ずしも参照先の意味的内容を表しているとは限らない。

【００１５】

以上の問題を鑑み、用途にあった文書を言語に依存せず、かつ精度良く迅速に収集することを可能とすることが、本発明が解決しようとする課題である。

【００１６】

【課題を解決するための手段】

本発明は、ネットワークから文書の収集を行なう装置または方法を前提とする。そして、本発明の各態様に係わる装置では、ネットワークから文書を収集する文書収集装置において、収集済みの文書群の参照関係に基づいて、次に収集すべき文書の候補である次収集候補を決定する次候補判定手段と、ネットワークから前記次収集候補を収集して収集済み文書群に加える文書収集手段と、を備え、収集済み文書群の文書がある数以上になるまで、次候補判定手段による次収集候補の決定及び前記文書収集手段による文書の収集を繰り返す。

【００１７】

上記装置を、ネットワーク上のコミュニティにとって有用度の高い文書を収集するコミュニティ向けの文書収集装置として構成するようにしてもよい。そのために、上記構成において、文書収集手段がネットワーク上のコミュニティ内から文書をまんべんなく収集した後、次候補判定手段は、収集済み文書群の参照関係に基づいてコミュニティ内外の文書から次収集候補を決定する、こととしてもよい。

【００１８】

コミュニティ内外から文書を収集する前に、コミュニティ内から文書をまんべんなく収集することにより、コミュニティ内で必要とされている多様な分野の文書についての情報を入手することができる。このようにして入手した多様な分野に関する文書群の参照関係を用いてコミュニティ内外から文書を収集することにより、正確にコミュニティにとって有用度の高い文書を収集することが可能となる。また、文書本文の内容を解析しないため、言語に依存せず、迅速にコミュニティにとって有用度の高い文書を収集することが可能となる。

【００１９】

上記構成において、収集済み文書群の参照関係及び文書のネットワーク上の場所を示す情報、例えばＵＲＬ、に基づいて重要度を算出するランキング手段を更

に備え、次候補判定手段は、参照関係及び重要度に基づいて次収集候補を決定することとしてもよい。

【００２０】

上記コミュニティ向け文書収集装置において、ランキング手段は、重要度に基づいて、前記コミュニティ内外に分けてランキングし、次候補判定手段は、コミュニティ内及びコミュニティ外それぞれにおいて、ランキングが高い文書を前記次収集候補とすることとしてもよい。これにより、次収集候補がコミュニティ内又はコミュニティ外に集中し、文書がコミュニティ内又はコミュニティ外いずれかからばかり収集されてしまうことを防ぐことが可能となる。

【００２１】

また、上記コミュニティ向け文書収集装置は、更に、収集済み文書群を検索した結果を、前記コミュニティ内外に分けて提示する提示手段を備えることとしても良い。これにより、コミュニティに属するクライアントが、コミュニティ内外別に文書の検索結果を取得することが可能となる。

【００２２】

また、上記コミュニティ向け文書収集装置は、更に、文書がコミュニティ内の文書であるか否かを文書のネットワーク上での場所を示す情報、例えばＵＲＬ、に基づいて判別するコミュニティ判別手段を備えることとしても良い。文書のネットワーク上での場所を示す情報に基づいて判定することにより、文書がコミュニティ内の文書であるか否かの判定が迅速に行うことが可能となる。

【００２３】

また、上記のネットワークから文書を収集する文書収集装置を、特定の分野に関する文書を収集する特定分野向け文書収集装置として構成するようにしてもよい。そのために、本発明の更なる別の態様によれば、ネットワークから文書を収集する装置において、文書の収集に先立って、特定分野に関する文書群である正例文書群と、特定分野と関連が少ない分野に関する文書群である負例文書群とを収集済み文書群として与え、文書収集手段は、収集された次収集候補を、正例文書群に加え、収集済み文書群のうち、正例文書群の文書がある数以上になるまで、次候補判定手段による次収集候補の決定及び文書収集手段による収集を繰り返

すように構成する。これにより、特定分野に関する文書を、文書本文の内容を解析せずに、参照関係に基づいて迅速に収集することが可能となる。

【００２４】

また、上記の特定分野向け文書収集装置において、更に、収集済み文書の参照関係に基づいて、正例文書群の文書からのみ参照される度合いである参照度を算出する参照度算出手段を備え、次候補判定手段は、参照度が高い文書を次収集候補として決定することとしてもよい。また、上記の特定分野向け文書収集装置において、更に、収集済み文書の参照関係に基づいて、正例文書群の文書を参照している収集済み文書群から参照されている文書について、収集済み文書群からの被参照数を示す共参照度を算出する共参照度算出手段を備え、次候補判定手段は、共参照度が高い文書を次収集候補として決定することとしてもよい。参照度及び共参照度を用いることにより、収集したい分野に関する文書を、文書本文の内容を検討すること無く、迅速に収集することが可能となる。

【００２５】

また、上記の特定分野向け文書収集装置は、複数の分野を対象とし、各分野に関する文書を同時に収集する文書収集装置とすることもできる。そのために、上記の特定分野向け文書収集装置において、収集に先立って与える収集済み文書群を複数の分野に関する文書群の和集合とし、ある分野に関する文書群を正例文書群として文書を収集する際に、他の残りの分野に関する文書群の和集合を負例文書群とするように構成する。

【００２６】

また、各文書収集装置において、更に、収集済み文書で用いられている参照表現に基づいて収集済み文書群をまとめあげるまとめあげる手段を更に備えることとしてもよい。参照表現のうち、参照先文書と参照元文書の内容が同一であるのにネットワーク上で分散されて格納されていることを示す参照表現がある。例えば、「次へ」、「Ｎｅｘｔ」、「前へ」及び「Ｐｒｅｖ」等がそのような参照表現に該当する。まとめあげ手段は、このような参照表現による参照関係をもつ２つ以上の文書を１つにまとめあげる。

【００２７】

また、各文書収集装置において、更に、収集済み文書群の文書である収集済み文書で用いられている参照表現に基づいて、収集済み文書にキーワードを付与するキーワード付与手段を備えることとしても良い。これにより、文書本文の意味内容を解析することなく、かつ、様々な各キーワードの異称をも、キーワードとすることが可能となる。

【００２８】

また、キーワード付与手段は、参照表現が参照先文書に関係なく使用される参照表現の場合、キーワードとしないこととしても良い。ここで、参照先文書に関係なく使用される参照表現の例として、「トップへ戻る」、「ホームへ」等が考えられる。

【００２９】

また、キーワード付与手段は、参照表現が参照する相異なる文書数を計数し、相異なる文書数がある数以上である場合、その参照表現をキーワードとしないこととしても良い。このような参照表現は、参照先文書に関係なく使用される参照表現である可能性が高いからである。

【００３０】

また、キーワード付与手段は、参照表現が参照する相異なる文書数がある数未満である場合、更に、各収集済み文書でその参照表現が参照されている回数である参照回数を計数し、相異なる文書数及び参照回数に基づいて、その参照表現をキーワードとするか否か判定することとしてもよい。

【００３１】

また、キーワード付与手段は、参照表現に基づくキーワードに、収集済み文書の本文から抽出したキーワード及び収集済み文書のＵＲＬから抽出したキーワードを組み合せることとしてもよい。これにより、多様な方法で抽出したキーワードを組み合せることが可能となる。

【００３２】

また、本発明の各構成により行われる処理の過程からなる方法によっても、前述した課題を解決することができる。また、上述した本発明の各構成により行な

われる機能と同様の制御をコンピュータに行なわせるプログラムを記憶したコンピュータ読み取り可能な記憶媒体から、そのプログラムをコンピュータに読み出させて実行させることによっても、前述した課題を解決することができる。

【００３３】

【発明の実施の形態】

以下、本発明の実施の形態を図面に基づいて説明する。本発明は、ネットワークから、用途にあった文書を収集する文書収集装置に関する。図１に、本発明の原理図を示す。図１に示すように、文書収集装置１は、インターネットやイントラネット等のネットワークに接続されている。文書収集装置１は、文書収集手段２、参照関係抽出手段３、コミュニティ判別手段４、次候補判定手段５、ランキング手段６、ＵＲＬ判定手段７、参照度／共参照度算出手段８、まとめあげ手段９、キーワード抽出手段１０を備える。図１において、点線で示される構成要素、つまり、コミュニティ判別手段４及び参照度／共参照度算出手段８は、実施形態によって用いられたり、用いられなかったりする。同様に、点線で示される矢印、つまり、ランキング手段６による文書のランキング結果は、実施形態によって、次候補判定手段１５による次収集候補の判定に用いられたり、用いられなかったりする。

【００３４】

本発明の１実施形態に係わる文書収集装置は、ネットワーク上のコミュニティ向けの文書を収集する。そのために、１実施形態に係わるコミュニティ向け文書収集装置は、文書収集手段２、参照関係抽出手段３、コミュニティ判別手段４、次候補判定手段５、ランキング手段６、まとめあげ手段９及びキーワード付与手段１０を備える。コミュニティ向け文書収集装置において、まず、コミュニティ内からまんべんなく文書を収集した後、コミュニティ内外からコミュニティにとって有用度が高い文書を収集する。

【００３５】

参照関係抽出手段３は、収集済み文書群２０から参照関係を抽出し、文書間参照関係２２を抽出する。なお、収集開始時は、予め収集済み文書群２０として初期文書群を与える。コミュニティ判別手段４は、収集済み文書群２０の参照先文

書であって、未収集の文書がコミュニティ内の文書であるか否か判別する。

【００３６】

次候補判定手段５は、収集済み文書群２０の参照先であって、コミュニティ内の未収集文書を次収集候補２１として判定する。文書収集手段２は、次収集候補２１として判定された文書を収集し、新たに収集した文書群（新規収集文書群）を収集済み文書群２０に加え、新たな収集済み文書群２０とする。文書収集手段２は、収集済み文書群２０の文書数が規定された値以上であるか否か判定する。収集済み文書群２０の文書数が規定された値より少ない場合、上述のようにしてコミュニティ内から文書を収集する処理を繰り返す。このようにコミュニティ内の文書を規定数以上、まんべんなく収集することにより、コミュニティ内の文書が属する多様な分野についての情報を取得する。この情報は、コミュニティにとって有用度が高い文書をコミュニティ内外から収集することに役立てられる。

【００３７】

収集済み文書群２０の文書数が規定された値以上である場合、次にコミュニティにとって有用度が高い文書をコミュニティ内外から収集する。参照関係抽出手段３により新規収集文書群から参照関係を抽出し、コミュニティ判別手段４により参照先文書であって未収集の文書がコミュニティ内の文書であるか否か判別する。ランキング手段６は、参照関係及び、文書のネットワーク上での場所を示す情報、例えばＵＲＬ、の特徴に基づいて、収集済み文書の参照先となっている未収集の文書をコミュニティ内外別にランキングする。ランキング手段６は、ＵＲＬ判定手段７を備え、ＵＲＬ判定手段７は、参照先文書と参照元文書のＵＲＬ文字列上の類似を判定する。ランキング手段６は、ＵＲＬ判定手段７によって判定されＵＲＬの文字列上の類似に基づいて、文書をランキングする。

【００３８】

次候補判定手段５は、コミュニティ内外でそれぞれ上位にランキングされた未収集文書を次回にネットワークから収集すべき文書である次収集候補２１として判定し、文書収集手段２は、次収集候補２１として判定された文書を収集する。このように、本発明の１実施形態に係わるコミュニティ向け文書収集装置は、多段階に分けてコミュニティにとって有用度が高い文書を収集する。ある規定され

た以上の文書をコミュニティ内外から収集すると、まとめあげ手段９は、参照表現に基づいて収集済み文書２０をまとめあげる。キーワード付与手段１０は、参照表現及び参照表現の出現頻度に基づいて、収集済み文書２０にキーワードを付与する。ランキング手段６は、上述のようにして、今度は収集済み文書２０をランキングする。最終的にまとめあげられ、キーワードを付与し、ランキングした収収集済み文書２０は、収集文書ファイル２３として格納される。上述のように、コミュニティ向け文書収集装置において、文書本文の内容を解析していないため、言語に依存せず、迅速に、用途に合った文書を収集することができる。

【００３９】

また、本発明の別の１実施形態に係わる文書収集装置は、特定の分野に関する文書を収集する。そのために、上記特定分野に関する文書収集装置は、文書収集手段２、参照関係抽出手段３、次候補判定手段５、ランキング手段６、参照度／共参照度算出手段８、まとめあげ手段９及びキーワード付与手段１０を備える。特定分野に関する文書収集装置において、コミュニティ内外の文書の区別は不要であるため、コミュニティ判別に係わる処理はない。

【００４０】

特定分野に関する文書収集装置において、収集に先立って特定分野に関する文書群を正例文書群として、その特定分野との関連が少ない文書群を負例文書群として与える。収集済み文書群２０は、正例文書群と負例文書群の和集合とする。参照度／共参照度算出手段８は、ある文書と正例文書群、その文書と負例文書群のそれぞれの参照関係に基づいて、その文書が特定分野に関連する度合いを参照度及び共参照度として算出する。次候補判定手段５は、ランキング手段６によるランキングの代わりに、参照度／共参照度算出手段８が算出した参照度又は共参照度が高い未収集文書を次収集候補として判定する。また、負例文書群に含まれる収集済み文書２０のうち、参照度又は共参照度が高い文書を負例文書群から除き、正例文書群に加える。文書収集手段２は、次収集候補２１として判定された文書を収集し、正例文書群に加える。そして、正例文書群の文書数が規定された数以上になるまで、次収集候補の決定及び文書の収集を繰り返す。その他の動作は、上述の通りである。

【００４１】

以下、第１実施形態に係わる、コミュニティにとって有用度の高い文書を収集するコミュニティ向け文書収集装置について説明する。本発明の第１実施形態において述べるネットワーク上のコミュニティとして、例えば、社内サイト、業界サイト及び特定トピックのネットワーク上のユーザグループが考えられる。ここで、社内サイトは、しばしばイントラネットに代表される。業界サイトは、複数の会社のシステムからなるエクストラネットに代表される。なお、社内サイトに必要な文書を収集する文書収集装置は、コーポレートポータル（ＥＩＰ：Enterprise Information Portal ともいわれる）ともいわれる、企業内のイントラネットポータルに適用可能である。

【００４２】

コミュニティのポータルにおいて、コミュニティにとって有用度が高い文書を優先して自動収集するという要件が必要とされている。例えば、コーポレートポータルの場合、業務に関係する文書を自動収集する必要がある。本発明の第１実施形態によれば、このような文書の自動収集を実現する。そのために、第１実施形態に係わる文書収集装置において、以下の考え方を採用する。

・特定のコミュニティにとって有用度の高い文書は、そのコミュニティの文書の多くからよく参照されている文書である、またはコミュニティの重要文書から参照されている文書である、と考える。

【００４３】

図２は、第１実施形態に係わる文書収集装置の構成を示す。図１に示すように、文書収集装置１００は、文書収集部１０１、参照関係抽出部１０２、コミュニティ判別部１０３、次候補判定部１０４、ランキング部１０５、まとめあげ部１０６及びキーワード付与部１０７を備える。

【００４４】

上述のように、本文書収集装置１００において、先にコミュニティ内の文書について複数回、収集を行い、次に、コミュニティ内外の文書についても複数回、収集を行う。このように多段階に分けて複数回、文書収集を行うことが本文書収集装置１００の特徴の１つである。

【００４５】

収集開始に先立って、まず、初期文書群を収集済み文書群Ｓとして与える。この初期文書群は、収集の開始点となる。初期文書群として、例えば、サイトのトップ文書やトップ文書の参照集等が考えられる。収集済み文書群Ｓ又は初期文書群は、具体的には、ＵＲＬテーブル１２０として文書収集装置１００に備えられる。

【００４６】

続いて、参照関係抽出部１０２は、収集済み文書群Ｓから参照関係を抽出し、収集済み文書群Ｓの参照先となる文書（以下、参照先文書という）のＵＲＬをＵＲＬテーブル１２０に格納し、抽出された参照関係を参照関係テーブル１２１に格納する。コミュニティ判別部１０３は、参照関係抽出部１０２が抽出した、収集済み文書群Ｓの参照先文書が、コミュニティ内の文書であるのか、コミュニティ外の文書であるのか、ＵＲＬに基づいて判定し、判別結果を参照関係テーブル１２１に格納する。

【００４７】

本文書収集装置１００は、先にコミュニティ内の文書について１回以上収集を行う。この際、収集をまんべんなく行う。次候補判定部１０４は、参照関係抽出部１０２が抽出した収集済み文書群Ｓの参照先文書のうち、まだ収集されていない、コミュニティ内の文書を次に収集すべき文書の候補（以下、次収集候補Ｎという）として判定する。文書収集部１０１は、次収集候補Ｎとして判定された文書群を収集し、収集した文書を収集済み文書群に追加し、新たな収集済み文書群Ｓとする。このコミュニティ内の文書の収集は、規定された数の文書を収集するまで行う。コミュニティ内の全ての文書を収集しなくても良く、大体、コミュニティ内の全文書の１／２から１／４程度で良い。まんべんなくコミュニティ内の文書を収集することにより、コミュニティ内で有用な文書の分野についての情報を入手する。

【００４８】

文書収集部１０１がコミュニティ内の文書を規定された数だけ収集した後、文書収集装置１００は、次に、コミュニティ内外の文書についても１回以上収集を

行う。この場合、上述のようにして、文書収集部１０１は、文書を収集し、参照関係抽出部１０２及びコミュニティ判別部１０３は、ＵＲＬテーブル１２０及び参照関係テーブル１２１に情報を格納した後、さらに、ランキング部１０５は、参照関係及び文書のＵＲＬに基づいて、参照先文書に重要度を与え、その重要度に基づいて、参照先文書をランキングする。

【００４９】

候補判定部１０４は、ランキング部１０５による判定結果に基づいて、まだ収集されていない参照先文書であって、コミュニティ内の文書のうちで上位ｎ１位内にある文書群、及び、コミュニティ外の文書のうちで上位ｎ２位内にある文書群を次収集候補Ｎとなる文書として判定する。コミュニティ内外で分けて次収集候補Ｎを決定することにより、コミュニティ内とコミュニティ外のいずれかに文書が偏って収集されてしまうことを防ぐことが可能となる。

【００５０】

続いて、コミュニティ内の文書の収集と同様にして、文書収集部１０１は、次収集候補Ｎをコミュニティ内外から収集し、収集した文書を収集済み文書群に追加して新たな収集済み文書群Ｓとする。文書収集装置１００は、規定された数の文書を収集するまで、コミュニティ内外からの文書収集を繰り返す。

【００５１】

文書収集部１０１がコミュニティ内外から規定数だけの文書を収集した後、収集した文書の選別を行う。文書の選別は、まとめあげ部１０６、キーワード付与部１０７及びランキング部１０５により行われる。まず、まとめあげ部１０６は、文書において他文書を参照する際に用いる文字列（参照表現ともいう）に基づいて、収集済み文書のうち、同一内容であるが複数の文書に分割されてい文書をまとめあげる。

【００５２】

キーワード付与部１０７は、文書中の参照表現に基づいて、キーワードを決定し、文書にキーワードを付与する。より具体的には、キーワード付与部１０７は、参照表現のうち、「トップに戻る」、「ホームへ」というような参照先に関係なくしばしば使用される参照表現を除く。続いて、キーワード付与部１０７は、

各参照表現が参照する相異なる文書数を計数し、参照表現テーブル１２２に格納うする（図２では不図示）。また、各参照表現について各収集済み文書での出現頻度を計数し、参照回数テーブル１２３に格納する（図２では不図示）。キーワード付与部１０７は、これら計数結果に基づいて各収集済み文書について参照表現の重みを算出し、重みが大きい順にある数だけの参照表現をキーワードとして各収集済み文書に付与する。

【００５３】

ランキング部１０５は、参照関係及び文書のＵＲＬに基づいて、各文書に重要度を付与し、その重要度に基づいて文書をランキングする。このように、本実施形態に係わるコミュニティ向け文書収集装置１００は、文書本文の内容を解析すること無く、参照関係及びＵＲＬに基づいて文書を収集し、まとめあげ、キーワードを付与し、ランキングする。

【００５４】

上述のようにして、文書収集装置１００は、まとめあげられ、キーワードが付与され、ランキングされた文書群を優良コンテンツ１３０として提供する。優良コンテンツ１３０は、検索エンジン１４０を介して索引１４１として提供されたり、検索エンジン１４０を介してサーバ１６０に提供されたり、分類エンジン１５０によってディレクトリ編集されてサーバ１６０に提供されたりする。サーバ１６０のクライアントは、サーバ１６０に提供された優良コンテンツ１３０を、ブラウザ１７０を介して閲覧することができる。

【００５５】

以下、図３から図６を用いて各テーブルのデータ構造について説明する。図３にＵＲＬテーブル１２０のデータ構造の一例を示す。図３に示すように、ＵＲＬテーブルは、各文書について文書を識別する文書ＩＤ（Identification information）、文書のＵＲＬ、収集済みであるか否かを示す収集済みフラグ、コミュニティ内の文書であるか否かを示すコミュニティフラグ及び文書の重要度を格納する。文書ＩＤ及びＵＲＬは、参照関係抽出部１０２が収集済み文書の参照先文書を抽出した際に格納される。収集済みフラグは、文書収集部１０１がその文書を収集した際に「オン（１）」にされる。コミュニティフラグは、コミュニティ判

別部１０３がその文書がコミュニティ内の文書であると判定した場合に「オン（１）」にされる。重要度は、ランキング部１０５が文書の参照関係及びＵＲＬの文字列上の特徴に基づいて算出し、格納する。

【００５６】

図４に参照関係テーブル１２１のデータ構造の一例を示す。図４に示すように、参照関係テーブル１２１は、文書の参照関係に関する情報を格納する。より具体的には、参照関係テーブル１２１は、参照元文書の文書ＩＤである参照元文書ＩＤ、参照元文書によって参照されるコミュニティ内の文書の文書ＩＤである参照先文書$ＩＤ_1$、及び、参照元文書によって参照されるコミュニティ外の文書の文書ＩＤである参照先文書$ＩＤ_2$を格納する。これら情報は、参照関係抽出部１０２によって格納される。

【００５７】

図５に参照表現テーブル１２２のデータ構造の一例を示す。図５に示すように、参照表現テーブル１２２は、収集済み文書で各参照表現が用いられる頻度に関する情報を格納する。より具体的には、参照表現テーブル１２２は、各参照表現について、参照表現を識別する表現ＩＤ、参照表現（文字列）、参照表現が参照する相異なる文書の数である頻度ＤＦ（ｗ）、及び、キーワードとして用いるべきか否かを示す要否フラグを格納する。これら情報は全て、キーワード付与部１０７によって格納される。

【００５８】

図６に参照回数テーブル１２３のデータ構造の一例を示す。図６に示すように、参照回数テーブル１２３は、各収集済み文書が各参照表現で参照されている回数を格納する。これら情報は全て、キーワード付与部１０７によって格納される。

【００５９】

以下、第１実施形態に係わる文書収集装置が実現する特定のコミュニティにとって有用度の高い文書を収集する方法について説明する。説明において以下の表記法を用いる。

・ＬＴ（Ｓ）は、文書群Ｓの参照先となる文書群を示す。

・Ｘ－Ｙは、集合Ｘと集合Ｙの差集合を示す。

【００６０】

最初に、図７を用いて特定のコミュニティ向けの文書を収集する処理の大まかな流れについて説明する。まず、収集開始時に、収集済み文書群Ｓの初期文書群（収集の開始点となる文書群）としてコミュニティ内の文書を与える。

【００６１】

参照関係抽出部１０２による参照関係の抽出結果及びコミュニティ判別部１０３による、参照先文書がコミュニティ内の文書であるか否かの判別結果に基づいて、候補判定部１０４は、次収集候補Ｎを抽出する（ステップＳ１）。次収集候補Ｎを抽出する処理について、詳しくは後述する。

【００６２】

続いて、文書収集部１０１は、ＵＲＬテーブル１２０に格納されたＵＲＬに基づいて、次収集候補Ｎを収集し（ステップＳ２）、収集された文書の収集済みフラグをオンにする。これにより、文書収集部１０１は、新たに収集された次収集候補Ｎを収集済み文書群Ｓに加える。つまり、式Ｓ∪Ｎで示される文書群を新たに収集済み文書群Ｓとする。

【００６３】

文書収集部１０１は、収集済み文書群Ｓに含まれる文書数が規定された文書数以上であるか否か判定する（ステップＳ３）。この判定は、ＵＲＬテーブル１２０に格納された収集済みフラグが「オン（１）」になっている文書の数を計数することにより行う。収集済み文書群Ｓに含まれる文書数が規定された文書数以上でない場合（ステップＳ３：Ｎｏ）、次候補判定部１０４は、再度次収集候補を決定し（ステップＳ４）、ステップＳ２に戻る。２回目以降の次収集候補の決定において、今回の収集で新たに収集した文書（以下、新規収集文書という）についての参照関係抽出部１０２による参照関係の抽出結果、及び、コミュニティ判別部１０３による新規収集文書の参照先文書がコミュニティ内の文書であるか否かの判別結果に基づいて、候補判定部１０４は、未収集の参照先文書のうちコミュニティ内の文書を次収集候補Ｎとして抽出する。ステップＳ４の処理は、ステップＳ１と同様であるため、ステップＳ１について、後述する際に一緒に説明す

る。

【００６４】

収集済み文書群Ｓに含まれる文書数が規定された文書数以上である場合（ステップＳ３：Ｙｅｓ）、今度は、候補判定部１０４は、コミュニティ内外の文書から次収集候補を決定する。そのために、まず、参照関係抽出部１０２は、新規収集文書の参照関係の抽出し、コミュニティ判別部１０３は、新規収集文書の参照先文書がコミュニティ内の文書であるか否かを判別する。その後、ランキング部１０５は、収集済み文書及びその参照先文書、つまりＳＵＬＴ（Ｓ）に対して重要度を付与し、重要度に基づいて、未収集の参照先文書、つまりＬＴ（Ｓ）－Ｓのランキングを行う（ステップＳ５）。このステップＳ５の処理について詳しくは後述する。

【００６５】

続いて、次候補判定部１０４は、ＬＴ（Ｓ）－Ｓのうち、コミュニティ内の文書群のランキングで上位ｎ１件に入っている文書群及びコミュニティ外の文書群のランキングで上位ｎ２件に入っている文書群を次収集候補Ｎとする（ステップＳ６）。このようにコミュニティ内とコミュニティ外とを区別して次収集候補Ｎを抽出することにより、コミュニティ内またはコミュニティ外に、収集される文書が偏ることを防ぐことができる。

【００６６】

文書収集部１０１は、ＵＲＬテーブル１２０に格納されたＵＲＬに基づいて、次収集候補Ｎを収集し（ステップＳ７）、収集された文書の収集済みフラグを「オン（１）」にする。文書収集部１０１は、ＵＲＬテーブル１２０に格納された収集済みフラグが「オン（１）」になっている文書の数を計数することにより、収集済み文書群Ｓに含まれる文書数が規定された文書数以上であるか否か判定する（ステップＳ８）。

【００６７】

収集済み文書群Ｓに含まれる文書数が規定された文書数以上でない場合（ステップＳ８：Ｎｏ）、ステップＳ５に戻る。収集済み文書群Ｓに含まれる文書数が規定された文書数以上である場合（ステップＳ８：Ｙｅｓ）、ランキング部１０

５、まとめあげ部１０６及びキーワード部１０７によって、収集済み文書群Ｓの文書を選別する（ステップＳ９）。ステップＳ９の処理について詳しくは後述する。

【００６８】

以下、コミュニティ内の文書を収集する際に、次収集候補を決定する処理について詳しく説明する。この処理は、図７のステップＳ１及びステップＳ４に相当する。

【００６９】

まず、参照関係抽出部１０２は、新規収集文書から参照されている参照先文書を抽出する（ステップＳ１１）。参照関係抽出部１０２は、各抽出された参照先文書について、参照先文書と同一のＵＲＬがＵＲＬテーブル１２０に格納されていない場合、参照先文書のＵＲＬをＵＲＬテーブル１２０に格納する（ステップＳ１２）。同じＵＲＬを重複して格納する必要はないからである。情報を格納する際、参照関係抽出部１０２は、収集済みフラグを「オフ（０）」とする。

【００７０】

続いて、コミュニティ判別部１０３は、ＵＲＬテーブル１２０に格納された参照先文書のＵＲＬの文字列に基づいて、抽出された参照先文書がコミュニティ内の文書であるか否か判別し、コミュニティ内の文書であると判別した場合、コミュニティ判別部１０３は、ＵＲＬテーブル１２０のコミュニティフラグを「オン（１）」とする。それ以外の場合、コミュニティ判別部１０３は、コミュニティフラグを「オフ（０）」とする（ステップＳ１３）。さらに、参照関係抽出部１０２は、コミュニティ判別部１０３の判別結果に基づいて、参照関係テーブル１２１の各欄に参照関係を格納する（ステップＳ１４）。

【００７１】

ここで、本実施形態によれば、コミュニティは、ネットワーク上の文書の集合、つまり文書群として与えられている。従って、同一コミュニティ内の文書であるか否かの判別は、その文書群を示すＵＲＬに基づいて判別できる。より具体的には、コミュニティ内の文書であるか否かの判定は、ＵＲＬの文字列上の特徴に基づいて、以下のようにして行う。

・コミュニティが社内サイトである場合、通常、社内サイトのドメイン名（fujitsu.co.jp等）とドメイン名が同じである文書をコミュニティ内の文書であると判定する。
・コミュニティが業界サイトである場合、その業界サイトに属する複数の企業のサイトのドメイン名のいずれかとドメイン名が同じである文書をコミュニティ内の文書であると判定する。
・コミュニティがユーザグループである場合、各ユーザのサイト（ホームページともいう）のＵＲＬ（例えば、http://www.fujitsu.co.jp/foo/ ）のいずれかと同じ文字列をＵＲＬに含む文書をコミュニティ内の文書であると判定する。

【００７２】

次候補判定部１０４は、収集済み文書の参照先文書であり、かつ、未収集文書である文書ＬＴ（Ｓ）－Ｓのうち、コミュニティ内の文書を次収集候補Ｎとして判定する。具体的には、次候補判定部１０４は、ＵＲＬテーブル１２０を参照し、収集済みフラグが「オフ（０）」であり、且つ、コミュニティフラグが「オン（１）」である文書を次収集候補Ｎとして決定する（ステップＳ１４）。このような次収集候補Ｎは、以下の（１）式で表すことができる。

【００７３】

Ｎ＝｛ｄ｜ｄ∈ＬＴ（Ｓ）－Ｓ，ｄはコミュニティ内｝　　・・・・（１）

このようにして次収集候補Ｎを決定し、コミュニティ内の文書をまんべんなく収集することにより、コミュニティ内で必要とされる、意味的に多様な文書についての情報を偏りなく取得することが可能となる。

【００７４】

続いて、図９を用いて収集済み文書及びその参照先文書をランキングする処理について説明する。この処理は、図７のステップＳ５に相当する。

参照関係抽出部１０２及びコミュニティ判別部１０３は、新規収集文書の参照関係の抽出し、参照関係をコミュニティの判別結果とともに、ＵＲＬテーブル１２０及び参照関係テーブル１２１に格納する（ステップＳ２１からＳ２３）。このステップＳ２１からＳ２３の処理は、図８で説明したステップＳ１１からＳ１３と同様であるため、詳しい説明は省略する。

【００７５】

続いて、ランキング部１０５は、収集済み文書及びその参照先文書、つまりＳＵＬＴ（Ｓ）に対して、参照関係テーブル１２１に格納された参照関係及びＵＲＬテーブル１２０に格納されたＵＲＬの文字列上の特徴に基づいて重要度を算出し、算出した重要度をＵＲＬテーブル１２０に格納する（ステップＳ２４）。ランキング部１０５は、ＵＲＬテーブル１２０に格納されたコミュニティフラグ及び重要度に基づいて、未収集の参照先文書、つまり、ＬＴ（Ｓ）－Ｓを、コミュニティ内外に分けてランキングする（ステップＳ２５）。

【００７６】

以下、ステップＳ２４の重要度を算出する処理について詳しく説明する。上述のように、ランキング部１０５は、文書の参照関係及びＵＲＬを利用して、収集済み文書の意味内容を分析することなく、文書の重要度を算出する。以下、参照関係に基づいて文書に付与される重要度をリンク重要度という。リンク重要度を付与する際の基本的な考え方は以下の通りである。

・類似度の低いＵＲＬから多く参照されている文書は重要である。

【００７７】

例えば、一般に、同一サイト内に設けられた複数の文書はそのサイト内の他の文書に参照されているが、それらの文書のＵＲＬは相互に類似する。従って、類似度の高いＵＲＬから参照されている文書の重要度は低いと推定できる。

・多くの文書から参照されている文書ほど重要な文書であり、重要な文書から参照されている、ＵＲＬの類似度の低い文書は重要である。

【００７８】

例えば、有名なディレクトリサービス等及び官公庁等は多くの文書から参照されているが、このような重要な文書から参照されている文書は重要度が高いと考えられる。また、多くの文書やミラーサイトを抱えるサービス（サイト）に設けられた文書等はそのサイト内で参照されていることが多いが、同じサイト内の文書のＵＲＬは大抵類似しているため、「ＵＲＬの類似度の低い文書は重要である」という考え方を導入すれば、同じサイトの文書が多く検索されてしまうことを避けることが可能となる。

・ＵＲＬの類似度は、サーバアドレス、パス、ファイル名の全てが異なるものが最も小さく、ミラーサイトや同一サーバ内の文書は類似度が高くなるように、ＵＲＬの字面情報から定義する。

【００７９】

上述の３つの考え方を導入することにより、全ての参照関係を同等に扱わないでリンク重要度に応じた重みを参照関係に与えることとしている。より具体的には、重みを参照元と参照先文書のＵＲＬの類似度の逆数として与えることとしている。以下、リンク重要度の算出についてより詳しく説明する。

【００８０】

リンク重要度の算出対象となる文書集合をＤＯＣ＝｛p_1，p_2，・・・，p_n｝、

文書ｐのリンク重要度をW_p、

文書ｐの参照先の文書集合をＲef(p)、

文書ｐの参照元の文書集合をＲefed(p)、

文書ｐとｑのＵＲＬ類似度をsim(p,q)、

相異度をdiff(p,q)＝1/sim(p,q)とすると、

文書ｐからｑに参照が張られているとした時、その参照の重みlw(p,q)を以下の（１）式で定義する。

【００８１】

【数１】

$$lw(p,q) = diff(p,q) / \sum_{i \in Ref(p)} diff(p,i)$$

$$= \frac{1}{sim(p,q) \sum_{i \in Ref(p)} \frac{1}{sim(p,i)}} \qquad \cdots\cdots\cdots(1)$$

【００８２】

この（１）式から分かるように、lw(p,q)は、ｐとｑのＵＲＬの類似度sim(p,q)が低いほど、また、ｐからの参照数がより少ないほど大きくなる。

各文書のリンク重要度は、各ｐ∈ＤＯＣに対して、C_q を定数（重要度の下限であり、文書によって異なる値を与えてもよい。）として、

【００８３】

【数２】

$$Wq = Cq + \sum_{p \in Refed(q)} W_p * Iw(p,q) \qquad \cdots\cdots\cdots(2)$$

【００８４】

という連立一次方程式の解として定義する。ランキング部１０５は、この連立一次方程式を解くことにより、リンク重要度を各文書に付与する。なお、連立一次方程式の解法については、既存のアルゴリズムが多数存在するため、説明は省略する。（１）式及び（２）式から、上述の考え方が実現されていることを読み取ることができる。

【００８５】

次に、（１）式及び（２）式中の文書ｐとｑのＵＲＬ類似度sim(p,q)について説明する。ＵＲＬ類似度は、ランキング部１０５のＵＲＬ判別部（不図示）により算出される。一般に、文書のＵＲＬは、サーバアドレス、パス、ファイル名の三種類の情報から構成される。例えば、ＷＷＷ文書のＵＲＬ、

http://www.flab.fujitsu.co.jp/hypertext/news/1999/product1.html は、サーバアドレス（www.flab.fujitsu.co.jp）、パス（hypertext/news/1999 ）、ファイル名（product1.html ）の３種類の情報から構成される。

【００８６】

本実施形態では、与えられた２つの文書ｐ及びｑのＵＲＬ類似度を、上記の三種類の組合せにより定義する。類似度sim(p,q)として、例えば、以下に述べるドメイン類似度sim __domain(p,q) 及び融合類似度sim __merge(p,q)が考えられる。

【００８７】

ドメイン類似度sim __domain(p,q) は、ドメインの類似に基づいて算出される

。ドメインとは、サーバアドレスの後半部分であり、会社や組織を表す。サーバアドレスが.com、.edu、.org等で終わる米国サーバの場合はサーバアドレスの後ろから２つめまで、サーバアドレスが.jp 、.fr 等で終わる他国のサーバの場合はサーバアドレスの後ろから３つめまでがドメインに相当する。

【００８８】

文書ｐと文書ｑのドメイン類似度は以下の式により定義される。

sim __domain(p,q) ＝１／α　　（ｐ、ｑが同一ドメインの場合）

＝１　　（ｐ、ｑが異なるドメインの場合）

ここで、αは定数で、０より大きく１より小さい実数値を取るとする。

【００８９】

また、sim(p,q)として、前述の三種類の情報を融合した類似度sim __merge(p,q)を次のように定義する。

sim __merge(p,q)＝（サーバアドレスの類似度）＋（パスの類似度）＋（ファイル名の類似度）

以下、右辺の各項の算出方法について説明する。

【００９０】

サーバアドレスの類似度は、アドレスの階層を後ろから見ていき、ｎレベルまで一致した場合、類似度を１＋ｎとする。例えば、www.fujitsu.co.jp とwww.flab.fujitsu.co.jpは３レベルまで一致しているので４となる。www.fujitsu.co.jp とwww.fujitsu.com は１レベルも一致していないので（一致０レベル）、類似度は１である。

【００９１】

パスの類似度は、先頭からパスの”/”で区切られた要素毎に比較し、一致したレベルまでを類似度とする。例えば、/doc/patent/index.htmlと/doc/patent/1999/2/file.htmlとは、２レベルまで一致しているので類似度は２である。

【００９２】

ファイル名の類似度は、ファイル名が一致する場合、類似度１とする。

このsim __merge(p,q)によっても、ＵＲＬが似通った文書が多く検索されることを防ぐことができる。

【００９３】

このようにして、ランキング部１０５は、文書に重要度を付与し、高い重要度を付与された文書を上位にランキングする。

このように、本実施形態によれば、ランキング部１０５は、取得した文書の参照関係及びＵＲＬの文字列の特徴に基づいて、文書本文の意味内容を解析せずに、つまり処理速度が速くかつ精度良く、文書に重要度を付与し、その重要度に基づいて文書をランキングすることができる。

【００９４】

以下、図１０を用いて収集済み文書を選別する処理について詳しく説明する。この処理は図７のステップＳ９に相当する。まず、まとめあげ部１０６は、収集済み文書群Ｓで用いられている参照表現に基づいて、収集済み文書群Ｓをまとめあげる（ステップＳ３１）。なお、参照表現とは、例えば、ＨＴＭＬ（Hyper Text Mark-up Language ）では、アンカータグで囲まれた部分がそれに相当する。

【００９５】

より具体的には、予め不図示のまとめあげ参照表現テーブルに、「次に」、「前へ」といった参照表現（参照時に用いられる文字列）を格納する。これら「次に」、「前へ」といった参照表現を用いている文書は、同一内容であるが、ＵＲＬが分散されている文書と推定される。まとめあげ部１０６は、まとめあげ参照表現テーブルに格納されている参照表現を文書から抽出し、以下のようにして文書をまとめあげる。

・文書ｄｏｃ１の中から「次へ」、「次に続く」、「Ｎｅｘｔ」というような表現により、文書ｄｏｃ２が参照されている場合、まとめあげ部１０６は、文書ｄｏｃ２を文書ｄｏｃ１に縮退する。この操作の繰り返しを可能な限り行う。

・文書ｄｏｃ１の中から「前へ」、「前に戻る」、「Ｐｒｅｖ」といった表現により、文書ｄｏｃ２が参照されている場合、まとめあげ部１０６は、文書ｄｏｃ１をｄｏｃ２に縮退する。この操作の繰り返しを可能な限り行う。

【００９６】

続いて、キーワード付与部１０７は、参照表現に基づいて収集済み文書Ｓにキーワードを付す（ステップＳ３２）。キーワード付与処理について詳しくは後述

する。最後に、ランキング部１０５は、上述の図９のステップＳ２４と同様にして、収集済み文書に重要度を付与し、重要度をＵＲＬテーブル１２０に格納する。ランキング部１０５は、重要度に基づいて収集済み文書をランキングする（ステップＳ３３）。

【００９７】

次に、ステップＳ３２のキーワード付与処理について、図１１を用いて詳しく説明する。まず、予め、収集済み文書で用いられている参照表現のうち、「ホームへ」、「トップに戻る」等、参照先文書に関係なく、しばしば使用される参照表現を不図示の不要語辞書に格納する（不図示）。キーワード付与部１０７は、収集済み文書群Ｓから参照表現を抽出し、各参照表現ｗについて、参照表現ｗを用いて参照される相異なる文書の数ＤＦ（ｗ）を集計し、参照表現ｗを識別する表現ＩＤ、その参照表現（文字列）とともにＤＦ（ｗ）の集計結果を参照表現テーブル１２２に格納する（ステップＳ４１）。この段階では、要否フラグを「オフ（０）」としておく。

【００９８】

キーワード付与部１０７は、参照表現ｗのうち、ＤＦ（ｗ）が所定の数以上であるものをキーワード候補から省く（ステップＳ４２）。言い換えると、参照先文書まで含めた総文書数をＮとすると、以下の（３）式に該当する参照表現ｗを省く。

【００９９】

$$DF(w) > \alpha N \quad \cdots\cdots (3)$$

ここで、αは、定数であり、例えば０．１としてもよい。

キーワード付与部１０７は、参照表現ｗのうち、不要語辞書に格納されている特定の参照表現をキーワード候補から省く（ステップＳ４３）。これらの参照表現は、参照先文書に関係なく使用されているため、キーワードとして用いるには適切でないからである。

【０１００】

キーワード付与部１０７は、収集済み文書Ｓから、文書ｄを取り出し、収集済み文書群ＳとｄのＳ－ｄを新たな収集済み文書群Ｓとする（ステ

ップＳ４４）。

【０１０１】

キーワード付与部１０７は、キーワード付与部１０７は、文書ｄで各参照表現ｗが参照されている回数ＴＦ（ｄ，ｗ）を集計し、以下の（４）式を用いて、文書ｄについて各参照表現ｗの重みＷ（ｄ，ｗ）を算出する（ステップＳ４５）。

【０１０２】

Ｗ（ｄ，ｗ）＝ＴＦ（ｄ，ｗ）ｌｏｇ（Ｎ／ＤＦ（ｗ））　・・・・（４）

キーワード付与部１０７は、参照表現テーブル１２２にアクセスし、参照表現の重みｗの大きい順に高々ｎ個の参照表現の要否フラグを「オン（１）」とする。つまり、重みｗの大きい順に高々ｎ個の参照表現を文書ｄのキーワードとする。

【０１０３】

このようにして得られた参照表現に基づくキーワードは、文書ｄの本文に含まれる単語に基づくキーワードと異なり、様々な異称をキーワードとして取得することが特徴の１つである。例えば、ある企業のホームページへの参照表現から、その企業の様々な呼称（正式名、略称、通称、英語名）を取得することができる。また、例えば、用語「Ｌｉｎｕｘ」に関して、「リナックス」、「ライナックス」等の様々内証がキーワードとして取得することができる。一方、一般に１つの文書の本文ではこうした異称のうち１つだけを統一的に用いるため、本文からキーワードを取得する場合では異称をキーワードとして取得することはできない。

【０１０４】

また、参照表現から取得したキーワードに、文書ｄの本文に出現する単語のうちで頻出する単語からキーワード及び文書ｄを示すＵＲＬから得たキーワード、例えば、http://www.fujitsu.com/ であれば、キーワードとしてfujitsu 、を加えることとしてもよい。これにより、文書ｄに多様なキーワードを付与することが可能になる。

【０１０５】

図１２に、第１実施形態に係わる文書収集装置を用いて収集した文書をユーザ

に提供する画面の一例を示す。図１２において、収集した優良コンテンツ１３０を、分類エンジン１５０を用いてディレクトリに分け、サーバ１６０のクライアントに提供する場合を例としている。クライアントは、画面１８０でキーワードを入力する、又は、カテゴリを選択することにより、閲覧したい文書へのリンクまたはリンク集を画面に表示させることができる。

【０１０６】

クライアントがキーワードを入力した場合、画面１８１に示すようにキーワードに基づいて検索された文書へのリンクが、重要度と共に表示される。本実施形態によれば、入力されたキーワードの異称も合わせて検索することが可能である。カテゴリを選択した場合、画面１８２に示すように選択されたカテゴリに関連する文書へのリンク集が表示される。

【０１０７】

ここで、画面１８１及び画面１８２に示すように、検索された文書を提示する際に、ＵＲＬテーブル１２０に格納されたコミュニティフラグに基づいて、文書をコミュニティ内外に分けて提示することとしても良い。

【０１０８】

以下、第２実施形態に係わる文書収集装置について説明する。第２実施形態に係わる文書収集装置は、特定分野に関する文書を収集する。本実施形態に係わる文書収集装置において以下の考え方を採用する。

・ネットワークにおいて、参照の親子／兄弟関係にある文書は、内容的に似通っている傾向にある。ある程度の文書群としばしば親子／兄弟関係にあるとされる文書は、元文書群と同じような分野の内容である可能性が高い。元の文書群からと親子／兄弟関係にある文書のうち参照度（親子関係）や共参照度（兄弟関係）の高い文書を収集し、元文書群に繰り込み、という操作を多段階に繰り返すことで、当該分野に関する文書を収集していくことができる。

【０１０９】

図１３に第２実施形態に係わる文書収集装置の構成を示す。図１３に示すように第２実施形態に係わる文書収集装置２００は、文書収集部１０１、参照関係抽出部１０２、候補判定部１０４、参照度／共参照度算出部２０１、ランキング部

１０５、まとめあげ部１０６及びキーワード付与部１０７を備える。参照度／共参照度算出部２０１は、文書の参照関係に基づいて、ある文書が特定分野に関連している度合いを算出する。その他の各部の機能は、第１実施形態で説明した通りである。

【０１１０】

第２実施形態に係わる文書収集装置において、収集開始に先立って、まず、当該分野の代表的な文書を既存の検索エンジンやリンク集を用いて収集し、正例文書群ＰＳとして与え、当該分野と重ならない任意の文やの文書も同様にして収集して負例文書群ＮＳとして与え、ＰＳ∪ＮＳを収集済み文書群Ｓとする。この収集済み文書群Ｓが収集の開始点となる。

【０１１１】

参照関係抽出部１０２は、収集済み文書群Ｓから参照関係を抽出し、収集済み文書群Ｓの参照先文書のＵＲＬをＵＲＬテーブル１２０に格納し、抽出された参照関係を参照関係テーブル１２１に格納する。ここで、第２実施形態に係わる文書収集装置において、ＵＲＬテーブル１２０に、コミュニティフラグの代わりに正例文書群ＰＳに含まれる文書であるか否かを示す正例フラグの欄を含む。正例フラグは、正例文書群ＰＳに含まれる文書である場合に「オン（１）」となる。また、参照関係テーブル１２１に参照関係を格納する際、コミュニティ内外で分けることは不要となる。

【０１１２】

参照度／共参照度算出部２０１は、参照関係抽出部１０２が抽出した参照関係に基づいて、正例文書群ＰＳ及び負例文書群ＮＳと収集済み文書Ｓの参照先文書との関係を示す参照度及び共参照度を算出する。次候補判定部１０４は、参照度／共参照度算出部２０１が算出した参照度及び共参照度に基づいて、収集済み文書群Ｓの参照先文書であって、正例文書群ＰＳに含まれない文書のなかから所定の条件を満たす文書を次収集候補Ｎとして判定する。次候補判定部１０４は、次収集候補Ｎのうち負例文書群ＮＳに含まれている文書を負例文書群ＮＳから除き、正例文書群ＰＳに加える。

【０１１３】

文書収集部１０１は、ＵＲＬテーブル１２０を参照し、次収集候補Ｎのうち未収集文書を収集し、収集した文書を正例文書群ＰＳに加える。第２実施形態に係わる文書収集装置２００は、正例文書群ＰＳの文書数が規定された数以上になるまで、上述のようにして収集済み文書Ｓの参照関係を抽出し、参照関係に基づいて次収集候補Ｎを決定し、次収集候補Ｎを収集する処理を繰り返す。

【０１１４】

収集済み文書Ｓが規定された数以上になると、まとめあげ部１０６は参照表現に基づいて収集済み文書群Ｓをまとめあげ、キーワード付与部１０７は参照表現が用いられる頻度等に基づいて収集済み文書群Ｓにキーワードを付す。ランキング部１０５は、参照関係及びＵＲＬの文字列上の特徴に基づいて各収集済み文書Ｓの重要度を算出し、重要度に基づいて収集済み文書Ｓをランキングする。これにより、分野別優良コンテンツ２１０を作成する。このように、第２実施形態に係わる文書収集装置によれば、文書本文の内容を解析せずに、特定分野に関する文書を収集し、まとめあげ、キーワードを付与することができる。

【０１１５】

分野別優良コンテンツ２１０は、検索エンジン１４０を介してサーバ１６０に提供される。サーバのクライアントはブラウザ１６０を用いて検索サービスの提供を受けることができる。

【０１１６】

以下、第２実施形態に係わる文書収集装置が実現する特定分野に関する文書収集方法について説明する。まず、用いる表記法について説明する。

・ＬＴ（Ｂ）は、文書群Ｂの参照先文書集合を示す。

・ＬＴ（ｐ）は、文書ｐの参照先文書集合を示す。

・ＬＳ（ｄ，Ｘ）＝｛ｃ∈Ｘ｜ｃ refers ｄ｝は、文書集合Ｘのうち文書ｄを参照している文書の集合を示す。

・ＬＳ（Ａ，Ｘ）＝｛ｃ∈Ｘ｜∃ｄ∈Ａ，ｃ refers ｄ｝は、文書集合Ｘのうち集合Ａ中の少なくとも１文書を参照している文書の集合を示す。

・ＣＣ（ｄ，Ａ，Ｘ）＝ＬＳ（ｄ，Ｘ）∩ＬＳ（Ａ，Ｘ）は、文書集合Ｘのうちで、文書ｄ、及び集合Ａの文書（少なくとも１文書）の両方を参照している文書

の集合を示す。

【０１１７】

図１４に、ＬＴ（Ｓ）、ＬＴ（ｐ）、ＬＳ（ｄ，Ｘ）及びＬＳ（Ａ，Ｘ）について、各集合が意味する文書の参照関係を示す。図１４において黒丸は文書を示し、矢印は参照関係を示し、矢印の元が参照元、矢印の先が参照先を示す。図１４に示すように、ＬＴ（Ｂ）とＬＳ（Ａ，Ｘ）及びＬＴ（ｐ）とＬＳ（ｄ，Ｘ）は、それぞれ矢印が逆になっている、つまり参照先文書と参照元文書が入れかわった関係にあることが分かる。また、図１５に、ＣＣ（ｄ，Ａ，Ｘ）が意味する文書の参照関係を示す。

【０１１８】

以下、図１６を用いて特定分野に関する文書を収集する処理について説明する。第２実施形態に係わる文書収集装置によれば、「ＸＭＬ」や「Ｌｉｎｕｘ」といった、特定分野（ジャンル）に関する意味的に類似した文書を優先的に収集する場合に、文書本文の内容を解析する処理を行わずに、参照関係に基づいて収集することが可能である。

【０１１９】

まず、当該分野に属する代表的な文書を、既存の検索エンジンやリンク集から探し出して収集し、正例文書群ＰＳとする。同様にして当該分野とは重ならない分野に属する文書を、探し出して収集し、負例文書群ＮＳとする。この正例文書群ＰＳと負例文書群ＮＳが初期文書群となる。そして、ＰＳ及びＮＳの文書のＵＲＬ、収集済みフラグ（全て「オン（１）」）、及び正例フラグ（正例文書の場合「オン（１）」）をＵＲＬテーブル１２０に格納する。正例文書群ＰＳと負例文書群ＮＳの和集合ＰＳ∪ＮＳを収集済み文書群Ｓとする（ステップＳ５１）。ここで、例えば、当該分野を「コンピュータ」であるとすると、当該分野と重ならない分野の例として、「手芸」、「料理」、「美容」等が考えられる。

【０１２０】

参照関係抽出部１０２は、収集開始時は初期の収集済み文書群Ｓ（初期文書群）から、それ以降は新規収集文書から参照関係を抽出し（ステップＳ５２）、参照先文書のＵＲＬをＵＲＬテーブル１２０に格納し、参照関係を参照関係テーブ

ル１２１に格納する。この処理は、第１実施形態と同様である。

【０１２１】

参照度／共参照度算出部２０１は、抽出された参照関係に基づいて、収集済み文書群Ｓの参照先文書から正例文書群ＰＳに含まれる文書を除いた文書集合Ｔ（Ｓ）＝ＬＴ（Ｓ）－ＰＳに含まれる文書ｄ∈Ｔ（Ｓ）について、以下の（５）式を用いて参照度R_{score}（ｄ，ＰＳ，Ｓ）を算出する。次候補判定部１０５は、参照度R_{score}（ｄ，ＰＳ，Ｓ）が上位ｎ１件に入っている文書群をＮ１とする。（ステップＳ５３）。なお、収集済み文書が正例文書群ＰＳに含まれるか否かは、ＵＲＬテーブル１２０の正例フラグを参照することにより判定できる。

【０１２２】

【数３】

$$R_{score}(d, PS, S)=\log(|LS(d, PS)|)\cdot\frac{|LS(d, PS)|}{|LS(d, S)|}$$

【０１２３】

（５）式の第１項は、文書ｄを参照している正例文書群ＰＳの文書数の対数を示す。また、（５）式の第２項は、文書ｄを参照している収集済み文書数に対する、文書ｄを参照している正例文書群ＰＳの文書数の割合を示す。従って、Ｓのうち正例文書群ＰＳからのみ多く参照されている文書ｄほど、R_{score}（ｄ，ＰＳ，Ｓ）が大きな値を取ることが分かる。

【０１２４】

つまり、次候補判定部１０５は、参照度R_{score}（ｄ，ＰＳ，Ｓ）に基づいて、新規収集文書の参照先文書のうち、特定分野に関係ある正例文書群ＰＳから多く参照され、特定分野とあまり関係ない負例文書群ＮＳから参照されていない文書をＮ１として決定する。図１７に、文書ｄについて参照度を算出する際に、（６）式に含まれる各集合が意味する参照関係を示す。

【０１２５】

続いて、参照度／共参照度算出部２０１は、文書ｄ∈Ｔ（Ｓ）－Ｎ１について

、以下の（６）式を用いて共参照度C_{score}（ｄ，ＰＳ，Ｓ）を算出する。次候補判定部１０５は、ｄ∈Ｔ（Ｓ）－Ｎ１のうちで共参照度C_{score}（ｄ，ＰＳ，Ｓ）が上位ｎ２件に入っている文書群をＮ２とする（ステップＳ５４）。

【０１２６】

【数４】

$$C_{score}(d, PS, S) = \log\left(\sum_{p \in CC(d, s)} |LT(p) \cap PS|\right) \frac{|CC(d, PS, S)|}{|LS(d, S)|}$$

【０１２７】

（６）式の第１項の対数の中身は、文書ｄ及び正例文書群ＰＳの文書の両方を参照している収集済み文書ｐ全てについての、文書ｐの参照先文書であって正例文書群ＰＳに含まれる文書数の積和を示す。従って、共参照度C_{score}（ｄ，ＰＳ，Ｓ）は、文書ｄ及び正例文書群ＰＳの文書の両方を参照している収集済み文書ｐの数が多い文書ｄほど、及び、このような文書ｐの参照先文書であって正例文書群ＰＳに含まれる文書の数が多いような文書ｄほど、大きな値を取ることが分かる。言い換えると、正例文書群ＰＳの文書を参照している収集済み文書から参照されている文書ｄについて、その文書ｄを参照している収集済み文書の数が多い文書ｄほど、共参照度C_{score}（ｄ，ＰＳ，Ｓ）は、大きな値を取る。

【０１２８】

（６）式の第２項は、文書ｄの参照元となっている収集済み文書の数に対する、文書ｄと共に参照されている文書ｐの数の割合を示す。共参照度C_{score}（ｄ，ＰＳ，Ｓ）は、この割合が大きいほど大きな値を取る。図１８に、文書ｄについて共参照度を算出する際に、（６）式に含まれる各集合が意味する参照関係を示す。

【０１２９】

次候補判定部１０５は次収集候補Ｎ＝Ｎ１∪Ｎ２とする（ステップＳ５５）。次候補判定部１０５は、次収集候補ＮのＵＲＬをキーとしてＵＲＬテーブル１２０を検索し、次収集候補Ｎの正例フラグを「オン（１）」する。この処理により

、負例文書群ＮＳに含まれていたが、次収集候補として判定された文書が、負例文書群ＮＳから除かれ、正例文書群ＰＳに加えられることとなる（ステップＳ５６）。

【０１３０】

文書収集部１０１は、ＵＲＬテーブル１２０に格納されたＵＲＬに基づいて、次収集候補Ｎのうち未収集文書をネットワークから収集し、収集した文書に対応する収集済みフラグを「オン（１）」にする（ステップＳ５７）。この処理により、新規収集文書を正例文書群ＰＳに加える。文書収集部１０１は、ＵＲＬテーブル１２０を参照し、正例文書群ＰＳの文書数が規定された数以上であるか否か判定する（ステップＳ５８）。正例文書群ＰＳの文書数が規定された数以上でない場合（ステップＳ５８：Ｎｏ）、ステップＳ５２に戻って処理を繰り返す。

【０１３１】

正例文書群ＰＳの文書数が規定された数以上である場合（ステップＳ５８：Ｙｅｓ）、正例文書群ＰＳの文書を選別し（ステップＳ５９）、処理を終了する。文書の選別処理は、第１実施形態と同様であるため説明を省略する。

【０１３２】

このようにして、本実施形態によれば、文書本文の内容を解析することなく、特定分野に関する文書を精度よく、かつ迅速に収集することが可能となる。

以下、第２実施形態の変形例について説明する。負例文書群ＮＳは、集めることも難しいため、収集処理の後に廃棄することをさけて、有効利用することが望ましい。そこで、第２実施形態の変形例に係わる文書収集装置によれば、上記処理で収集した負例文書群ＮＳを有効に利用することとする。これにより、なるべく独立な、例えば、「Ｊａｖａ言語」と「編物」及び「フランス料理」等、複数分野の文書を並行して収集することを可能とする。そのために、ある分野の文書を収集する際、その分野の文書群を正例文書群とし、その分野以外の他の分野の文書群を負例文書群ＮＳとして扱う。

【０１３３】

文書収集装置の構成は、図１３を用いて説明した通りであるため、説明を省略する。以下、図１９を用いて第２実施形態の変形例に係わる文書収集装置で行う

処理について説明する。

【０１３４】

まず、ｎ個の独立な分野の文書群D_i（ｉ＝１，２，・・・ｎ）を、検索エンジンやリンク集等から探し出して収集し、文書群D_iの文書のＵＲＬ、収集済みフラグ、及び分野を識別する情報である分野識別情報をＵＲＬテーブル１２０に格納する。第２実施形態に係わる文書収集装置では、正例フラグは不要である。文書群D_iは、分野ｉの初期文書群となる。収集済み文書群をＤ＝（D_1，D_2，・・・，D_n）とする（ステップＳ６１）。

【０１３５】

まず、参照関係抽出部２０１は、ｉを与える（ステップＳ６２）。なお、収集開始時に、参照関係抽出部２０１は、ｉを１とする。続いて、参照関係抽出部２０１は、ｉがｎを超えているか否か判定する（ステップＳ６３）。ｉがｎを超えている場合（ステップＳ６３：Ｙｅｓ）、ステップＳ７１に進む。そうでない場合（ステップＳ６３：Ｎｏ）、参照関係抽出部１０２は、分野ｉに対応する文書群D_iの新規収集文書から（収集開始時は初期文書群から）、参照関係を抽出し、参照先文書のＵＲＬをＵＲＬテーブル１２０に、参照関係を参照関係テーブル１２１にそれぞれ格納する（ステップＳ６４）。この処理は、第１実施形態と同様である。

【０１３６】

参照度／共参照度算出部２０１は、文書群D_iの参照先文書であって、収集済み文書群Ｄに含まれない文書群Ｔ（D_i）＝ＬＴ（D_i）－Ｄを次収集範囲とし、この次収集範囲Ｔ（D_i）に含まれる文書ｄ∈Ｔ（D_i）について、上述の（５）式を用いて参照度R_{score}（ｄ，D_i，Ｄ）を算出する。次候補判定部１０５は、参照度R_{score}（ｄ，D_i，Ｄ）が上位ｎ１件に入っている文書群を$N1_i$とする。（ステップＳ６５）。なお、収集済み文書が含まれる分野は、ＵＲＬテーブル１２０の分野識別情報を参照することにより判定できる。

【０１３７】

参照度／共参照度算出部２０１は、次収集範囲Ｔ（D_i）から$N1_i$を除いた集合に含まれる文書ｄ∈Ｔ（D_i）－$N1_i$ついて、上述の（６）式を用いて共

参照度C_{score}（d，D_i，D）を算出する。次候補判定部１０５は、共参照度C_{score}（d，D_i，D）が上位ｎ２件に入っている文書群を$N2_i$とする。（ステップＳ６６）。

【０１３８】

次候補判定部１０５は、$N1_i \cup N2_i$を分野ｉについての次収集候補N_iとする（ステップＳ６７）。次候補判定部１０５は、ＵＲＬテーブル１２０にアクセスし、次収集候補N_iに現在のｉの値に対応した分類識別情報を付す。文書収集部１０１は、ネットワークから次収集候補N_iを収集する（ステップＳ６８）。文書収集部１０１は、ＵＲＬテーブル１２０にアクセスし、収集された次収集候補N_i（新規収集文書群）の収集済みフラグを「オン（１）」とする。これにより、文書収集部１０１は、文書群D_iに新規収集文書群を加えて新たな文書群D_iとする（ステップＳ６９）。

【０１３９】

続いて、参照関係抽出部１０２は、ｉを１インクリメントし（ステップＳ７０）、ステップＳ６３に戻る。文書収集装置２００は、上述の処理をｉがｎを超えるまで、処理を繰り返す。

【０１４０】

ｉがｎを超えると（ステップＳ６３：Ｙｅｓ）、参照関係抽出部１０２は、ＵＲＬテーブル１２０を参照し、収集済みフラグ及び分野識別情報に基づいて、各文書群D_iの文書数を計数し、各文書群D_iの文書数が規定された数以上であるか否か判定する（ステップＳ７１）。文書数が規定数以上でない文書群D_k（ｋは１からｎまでの任意の数）がある場合、ステップＳ６２に戻り、参照関係抽出部１０２は、ｉ＝ｋとしてステップＳ６３以下の処理を繰り返す。

【０１４１】

なお、文書数が規定数以上でない文書群D_kが複数ある場合、例えば、D_{k1}，D_{k2}及びD_{k3}がある場合、ｉ＝ｋ１，ｋ２及びｋ３である場合について、ステップＳ６３以下の処理を繰り返す。D_1からD_nまで全ての収集済み文書群D_iについて文書数が規定数以上である場合（ステップＳ７１：Ｙｅｓ）、処理を終了する。

【０１４２】

これにより、ある分野の文書を収集する際に、その分野の文書群を正例文書群ＰＳとし、他の残りの分野の文書群の和集合を負例文書群ＮＳとして用いることができるため、負例文書群ＮＳに関する処理が無駄にならないこととなる。

【０１４３】

また、第２実施形態の変形例によれば、ある分野の文書群D_iを正例文書群ＰＳとして、その分野に関する文書を収集する場合に注目すると、負例文書群ＮＳとして用いられる他の分野の文書群が、正例文書群ＰＳと比べ大きくなる。さらにまた、負例文書群ＮＳ自体も他の分野に関する文書群であるため、意味的に一定している。変形例ではない第２実施形態においてある程度以上収集が進むと、正例文書群ＰＳが大きくなる一方で負例文書群ＮＳから正例文書群ＰＳに文書が移されることによって、例えば（５）式に示されるR_{score}（ｄ，ＰＳ，Ｓ）の第２項が大きくなっていく事態が生じうる。これによって、収集の精度が低下するる可能性があったが、変形例ではその可能性が低くなる。

【０１４４】

以下、図２０及び図２１を用いて、第２実施形態に係わる文書収集装置において特定分野に関する文書を収集する精度について説明する。図２０に、ネットワークから収集した約６７０万ＵＲＬの文書を全体集合Ｄとし、ＵＲＬに「Ｌｉｎｕｘ」を含む15,000ＵＲＬの文書を正解例Ｌとし、任意に選択した約5,000 ＵＲＬの文書を正例文書群ＰＳの初期文書とし、及びそれ以外の文書（Ｄ－ＰＳ）を負例文書群ＮＳの初期文書として、文書収集装置の収集精度を実験した結果を示す。

【０１４５】

図２０において、横軸に収集のくり返し回数ｉ、縦軸に適合率又は再現率を示す。再現率を折れ線、適合率を四角プロットで示す。ここで、ｉ回目の繰り返しで得られた正例集合PS_iについての適合率及び再現率は、以下（７）式及び（８）式で示される。

【０１４６】

$$適合率 = |PS_i \cap L| / |PS_i| \quad \cdots\cdots (7)$$

再現率＝｜ＰＳ$_i$∩Ｌ｜／｜Ｌ｜　・・・・（８）

つまり、適合率は、正例集合Ｓ$_i$中の正例文書群Ｓに含まれる正解例Ｌの割合であり、対象としている分野に含まれない文書（いわゆるゴミ）の少なさを示す。再現率は、正解例Ｌ中の正例文書群Ｓ$_i$に含まれる正解例Ｌの割合であり、対象としている分野に含まれる文書が収集されないこと（いわゆる漏れ）の少なさを示す。図２０に示すように、繰り返し回数が７３回程度になると、再現率が急激に低下するが、数十回の繰り返しでは、適合率、再現率とも良好であることが分かる。なお、繰り返し回数が７３回程度になると再現率が低下する原因は、所謂ゴミがゴミをよぶためであると考えられる。

【０１４７】

図２１に、ＵＲＬに「What's New」を含む14,000ＵＲＬを正解例Ｌとした場合に、同様の実験を行った結果を示す。図２１に示すように、繰り返し回数が数回程度になると急激に適合率が低下している。これは、What's Newのようなコンテンツは、互いにあまり意味的な関連（つながり）が無いためと考えられる。

【０１４８】

図２０に示す実験結果から、本実施形態に係わる文書収集装置によれば意味的に関連する文書群を効率よく収集することができることが分かる。

上述において説明した各サーバ及び各端末は、図２２に示すような情報処理装置（コンピュータ）を用いて構成することができる。図２２の情報処理装置２０は、ＣＰＵ３０１、メモリ３０２、入力装置３０３、出力装置３０４、外部記憶装置３０５、媒体駆動装置３０６、及びネットワーク接続装置３０７を備え、それらはバス３０８により互いに接続されている。

【０１４９】

メモリ３０２は、例えば、ＲＯＭ（Read Only Memory）、ＲＡＭ（Random Access Memory）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ３０１は、メモリ３０２を利用してプログラムを実行することにより、必要な処理を行う。

【０１５０】

上述の各サーバ及び各端末を構成する各機器及び各部は、それぞれメモリ３０

２の特定のプログラムコードセグメントにプログラムとして格納される。入力装置３０３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置３０４は、例えば、ディスプレイやプリンタ等であり、情報処理装置３００の利用者への問い合わせ、処理結果等の出力に用いられる。

【０１５１】

外部記憶装置３０５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。この外部記憶装置３０５に上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ３０２にロードして使用することもできる。

【０１５２】

媒体駆動装置３０６は、可搬記録媒体３０９を駆動し、その記録内容にアクセスする。可搬記録媒体３０９としては、メモリカード、メモリスティック、フロッピーディスク、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、光ディスク、光磁気ディスク、ＤＶＤ（Digital Versatile Disk）等、任意の情報処理装置で読み取り可能な記録媒体が用いられる。この可搬記録媒体３０９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ３０２にロードして使用することもできる。

【０１５３】

ネットワーク接続装置３０７は、ＬＡＮ、ＷＡＮ等の任意のネットワーク（回線）を介して外部の装置を通信し、通信に伴なうデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ３０２にロードして使用することもできる。

【０１５４】

図２３は、図２２の情報処理装置３００にプログラムとデータを供給することのできる情報処理装置で読み取り可能な記録媒体及び伝送信号を示している。

なお、本発明は、情報処理装置により使用されたときに、上述の本発明の実施形態の各構成によって実現される機能と同様の機能を情報処理装置に行わせるための情報処理装置で読み出し可能な記録媒体３０９として構成することもできる

【０１５５】

実施形態において各装置により行なわれる処理と同様のものを情報処理装置に行なわせるプログラムを、情報処理装置で読み取り可能な記録媒体３０９に予め記憶させておき、図２３に示すようにしてその記録媒体３０９からそのプログラムを情報処理装置３００に読み出させてその情報処理装置３００のメモリ３０２や外部記憶装置３０５に一旦格納させ、その情報処理装置３００の有するＣＰＵ３０１にこの格納されたプログラムを読み出させて実行させる。

【０１５６】

また、プログラム（データ）提供者３１０から情報処理装置３００にプログラムをダウンロードする際に回線３１１（伝送媒体）を介して伝送される伝送信号自体も、上述した本発明の実施形態において説明した各装置に相当する機能を汎用的な情報処理装置で行なわせることのできるものである。

【０１５７】

以上、本発明の実施形態について説明したが、本発明は上述した実施形態に限定されるものではなく、他の様々な変更が可能である。

例えば、第１実施形態に係わる文書収集装置１００と第２実施形態に係わる文書収集装置２００とを組みせるように構成ことにより、コミュニティ向けに分野別に文書を収集させることとしてもよい。

【０１５８】

また、文書収集装置１００又は２００を構成する各部及び各ＤＢは、お互いに連携して動作することにより一連のビジネスプロセスを実現する。これら各部及び各ＤＢは同じサーバに設けられてもよいし、異なるサーバに設けられネットワークを介して連携して動作することとしてもよい。

【０１５９】

（付記１）　ネットワークから文書を収集する文書収集方法であって、

前記文書の参照関係に基づいて、前記ネットワーク上のコミュニティ内から文書を所定数以上収集し、

前記コミュニティから前記所定数以上の文書を収集した後、収集済み文書の参

照関係に基づいて、前記コミュニティ内外から文書を収集する、

ことを特徴とする文書収集方法。

【０１６０】

（付記２）　前記収集済み文書群の参照関係及びネットワーク上の場所を示す情報に基づいて重要度を算出し、

前記参照関係及び前記重要度に基づいて、収集すべき文書を決定する、

ことを特徴とする付記１記載の文書収集方法。

【０１６１】

（付記３）　前記収集すべき文書は、前記コミュニティ内外別に決定される、

ことを特徴とする付記２記載の文書収集方法。

（付記４）　前記収集済み文書群を検索した結果を、前記コミュニティ内外に分けて提示する、

ことを特徴とする付記３記載の文書収集方法。

【０１６２】

（付記５）　前記コミュニティ内の文書であるか否かを前記ネットワーク上の場所を示す情報に基づいて判定する、

ことを特徴とする付記２記載の文書収集方法。

【０１６３】

（付記６）　ネットワークから文書を収集する文書収集方法であって、

ある分野に関する文書群である正例文書群と、前記分野と関連が少ない分野に関する文書群である負例文書群とを与え、

前記正例文書群及び前記負例文書群の参照関係に基づいて、前記分野に関する収集すべき文書を決定し、　前記ネットワークから前記収集すべき文書を収集する、

ことを特徴とする文書収集方法。

【０１６４】

（付記７）　前記参照関係に基づいて、前記正例文書群の文書からのみ参照される度合いである参照度を算出し、

前記参照度が高い文書を収集すべき文書として決定する、

ことを特徴とする付記６記載の文書収集方法。

【０１６５】

（付記８）　前記参照関係に基づいて、前記正例文書群の文書を参照している収集済み文書から参照されている文書について、収集済み文書からの被参照数を示す共参照度を算出し、

共参照度が高い文書を収集すべき文書として決定する、

ことを特徴とする付記６記載の文書収集方法。

【０１６６】

（付記９）　前記負例文書群は、複数の分野に関する文書群の和集合である、

ことを特徴とする、付記６記載の文書収集方法。

（付記１０）　前記収集済み文書で用いられている参照表現に基づいて、前記収集済み文書群をまとめあげる、

ことを特徴とする付記１記載の文書収集方法。

【０１６７】

（付記１１）　前記収集済み文書で用いられている参照表現に基づいて、前記収集済み文書にキーワードを付与する、

ことを特徴とする付記１記載の文書収集方法。

【０１６８】

（付記１２）　前記参照表現が参照先文書に関係なく使用される参照表現の場合、キーワードとしない、

ことを特徴とする付記１１記載の文書収集方法。

【０１６９】

（付記１３）　前記参照表現が参照する相異なる文書数を計数し、

前記相異なる文書数がある数以上である場合、前記参照表現をキーワードとしない、

ことを特徴とする付記１１記載の文書収集方法。

【０１７０】

（付記１４）　前記相異なる文書数がある数未満である場合、各収集済み文書で前記参照表現が参照されている回数である参照回数を計数し、

前記相異なる文書数及び前記参照回数に基づいて、前記参照表現をキーワードとするか否か判定する、

ことを特徴とする付記１１記載の文書集収集方法。

【０１７１】

（付記１５）　前記参照表現に基づくキーワードに、前記収集済み文書の本文から抽出したキーワード及び前記収集済み文書のＵＲＬから抽出したキーワードを組み合せる、

ことを特徴とする付記１１記載の文書集収集方法。

【０１７２】

（付記１６）　ネットワーク上のコミュニティに属する端末から文書を検索する検索方法であって、

文書を検索するための情報をサーバに送信し、

前記検索するための情報に基づいて前記コミュニティ内外に分けて検索した文書を、前記コミュニティにとっての有用度とともに受信する、

ことを特徴とする検索方法。

【０１７３】

（付記１７）　ネットワークから文書を収集する文書収集装置であって、

前記文書の参照関係に基づいて、次に収集すべき文書の候補である次収集候補を決定する次候補判定手段と、

前記文書のネットワーク上の場所を示す情報に基づいて前記文書が前記ネットワーク上のコミュニティ内の文書であるか否か判別するコミュニティ判別手段と、

前記ネットワークから前記次収集候補を収集する文書収集手段と、を備え、

前記文書収集手段は、前記コミュニティ内から所定数以上文書を収集した後、前記コミュニティ内外から文書を収集する、

ことを特徴とする文書収集装置。

【０１７４】

（付記１８）　ネットワークから文書を収集する文書収集装置であって、

ある分野に関する文書群である正例文書群及び前記分野と関連が少ない分野に

関する文書群である負例文書群の参照関係に基づいて、次に収集すべき文書の候補である次収集候補を決定する次候補判定手段と、

前記ネットワークから前記次収集候補を収集する文書収集手段とを備える、

ことを特徴とする文書収集装置。

【０１７５】

（付記１９）　コンピュータに実行させることによって、ネットワークから文書を収集する制御を該コンピュータに行なわせるプログラムを記録した、コンピュータで読み取り可能な記録媒体であって、

前記文書の参照関係に基づいて、前記ネットワーク上のコミュニティ内から文書を所定数以上収集し、　前記コミュニティから前記第１の所定数以上の文書を収集した後、収集済み文書の参照関係に基づいて、前記コミュニティ内外から文書を収集する、

ことを含む制御をコンピュータに行なわせるプログラムを記録した記録媒体。

【０１７６】

（付記２０）　コンピュータに実行させることによって、ネットワークから

コンピュータに実行させることによって、ネットワークから文書を収集する制御を該コンピュータに行なわせるプログラムを記録した、コンピュータで読み取り可能な記録媒体であって、

ある分野に関する文書群である正例文書群及び前記分野と関連が少ない分野に関する文書群である負例文書群の参照関係に基づいて、前記分野に関する収集すべき文書を決定し、

前記ネットワークから前記収集すべき文書を収集する、

ことを含む制御をコンピュータに行なわせるプログラムを記録した記録媒体。

【０１７７】

（付記２１）　搬送波に具現化された、ネットワークから文書を収集する制御をコンピュータに行わせるプログラムを表現するコンピュータ・データ・シグナルであって、前記プログラムは以下をコンピュータに実行させる、

前記文書の参照関係に基づいて、前記ネットワーク上のコミュニティ内から文書を所定数以上収集し、

前記コミュニティから前記所定数以上の文書を収集した後、収集済み文書の参照関係に基づいて、前記コミュニティ内外から文書を収集する、

【０１７８】

【発明の効果】

以上詳細に説明したように、本発明は、ある用途向けの文書を収集する際に、文書間の参照関係に基づいて収集すべき文書を決定し、決定された文書を収集することにより、言語に依存すること無く、迅速に用途にあった文書を選択して収集することが可能となる。

【０１７９】

また、参照表現に基づいて、収集済み文書をまとめあげ、各収集済み文書にキーワードを付与することにより、収集済み文書へのアクセスを容易とすることが可能となる。また、文書本文の内容を解析しないため、言語に依存せず、迅速にキーワードを付与することが可能となる。

【図面の簡単な説明】

【図１】

本発明の原理図である。

【図２】

第１実施形態に係わる文書収集装置の構成図である。

【図３】

ＵＲＬテーブルのデータ構造の一例を示す図である。

【図４】

参照関係テーブルのデータ構造の一例を示す図である。

【図５】

参照表現テーブルのデータ構造の一例を示す図である。

【図６】

参照回数テーブルのデータ構造の一例を示す図である。

【図７】

第１実施形態に係わる文書収集装置が行う処理の大まかな流れを示すフローチャートである。

【図８】

コミュニティ内の文書を収集する際に次収集候補を判定する処理を示すフローチャートである。

【図９】

収集済み文書及び参照先文書をランキングする処理を示すフローチャートである。

【図１０】

収集済み文書を選別する処理を示すフローチャートである。

【図１１】

キーワード付与処理を示すフローチャートである。

【図１２】

収集した文書を提供する画面の一例を示す図である。

【図１３】

第２実施形態に係わる文書収集装置の構成図である。

【図１４】

ＬＴ（Ｓ）、ＬＴ（ｐ）、ＬＳ（ｄ，Ｘ）、ＬＳ（Ａ，Ｘ）が意味する文書の参照関係を示す図である。

【図１５】

ＣＣ（ｄ，Ａ，Ｘ）が意味する文書の参照関係を示す図である。

【図１６】

第２実施形態に係わる文書収集装置が行う処理を示すフローチャートである。

【図１７】

参照度を算出する式に含まれる各集合が意味する参照関係を示す図である。

【図１８】

共参照度を算出する式に含まれる各集合が意味する参照関係を示す図である。

【図１９】

第２実施形態の変形例に係わる文書収集装置が行う処理を示すフローチャートである。

【図２０】

文書収集装置の収集精度の実験結果を示す図（その１）である。

【図２１】

文書収集装置の収集精度の実験結果を示す図（その２）である。

【図２２】

情報処理装置の構成図である。

【図２３】

情報処理装置にプログラムやデータを供給する記録媒体、伝送信号及び伝送媒体を説明する図である。

【符号の説明】

１、１００、２００　文書収集装置

２　文書収集手段

３　参照関係抽出手段

４　コミュニティ判別手段

５　次候補判定手段

６　ランキング手段

７　ＵＲＬ判定手段

８　参照度／共参照度算出手段

９　まとめあげ手段

１０　キーワード付与手段

２０　収集済み文書群

２１　次収集候補

２２　文書間参照関係

２３　収集文書ファイル

１０１　文書収集部

１０２　参照関係抽出部

１０３　コミュニティ判別部

１０４　候補判定部

１０５　ランキング部

１０６　まとめあげ部

１０７　キーワード付与部

１２０　ＵＲＬテーブル

１２１　参照関係テーブル

１２２　参照表現テーブル

１２３　参照回数テーブル

１３０　優良コンテンツ

１４０　検索エンジン

１４１　索引

１５０　分類エンジン

１６０　サーバ

１７０　ブラウザ

１８０、１８１、１８２　画面

２０１　参照度／共参照度テーブル

２１０　分野別優良コンテンツ

３００　情報処理装置

３０１　ＣＰＵ

３０２　メモリ

３０３　入力装置

３０４　出力装置

３０５　外部記憶装置

３０６　媒体駆動装置

３０７　ネットワーク接続装置

３０８　バス

３０９　可搬記録媒体

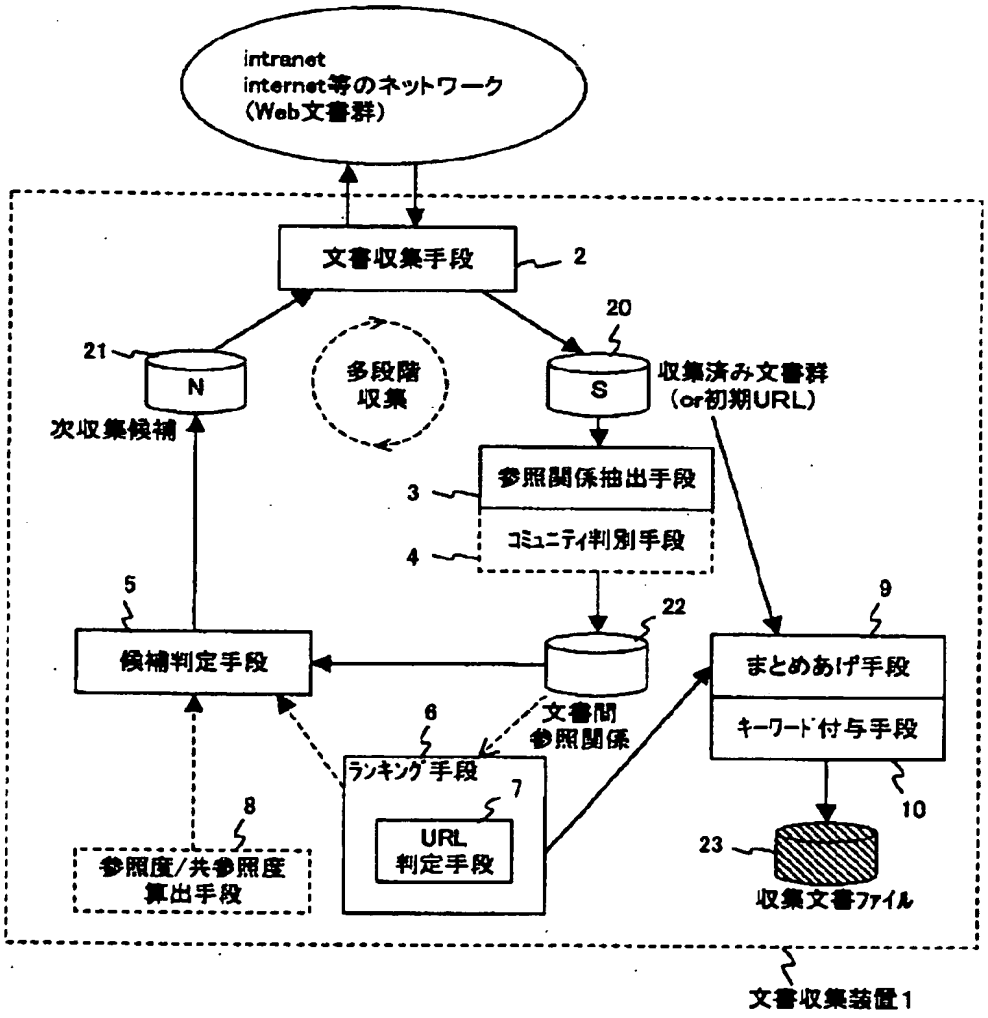
３１０　プログラム（データ）提供者

３１１　回線

【書類名】　　図面

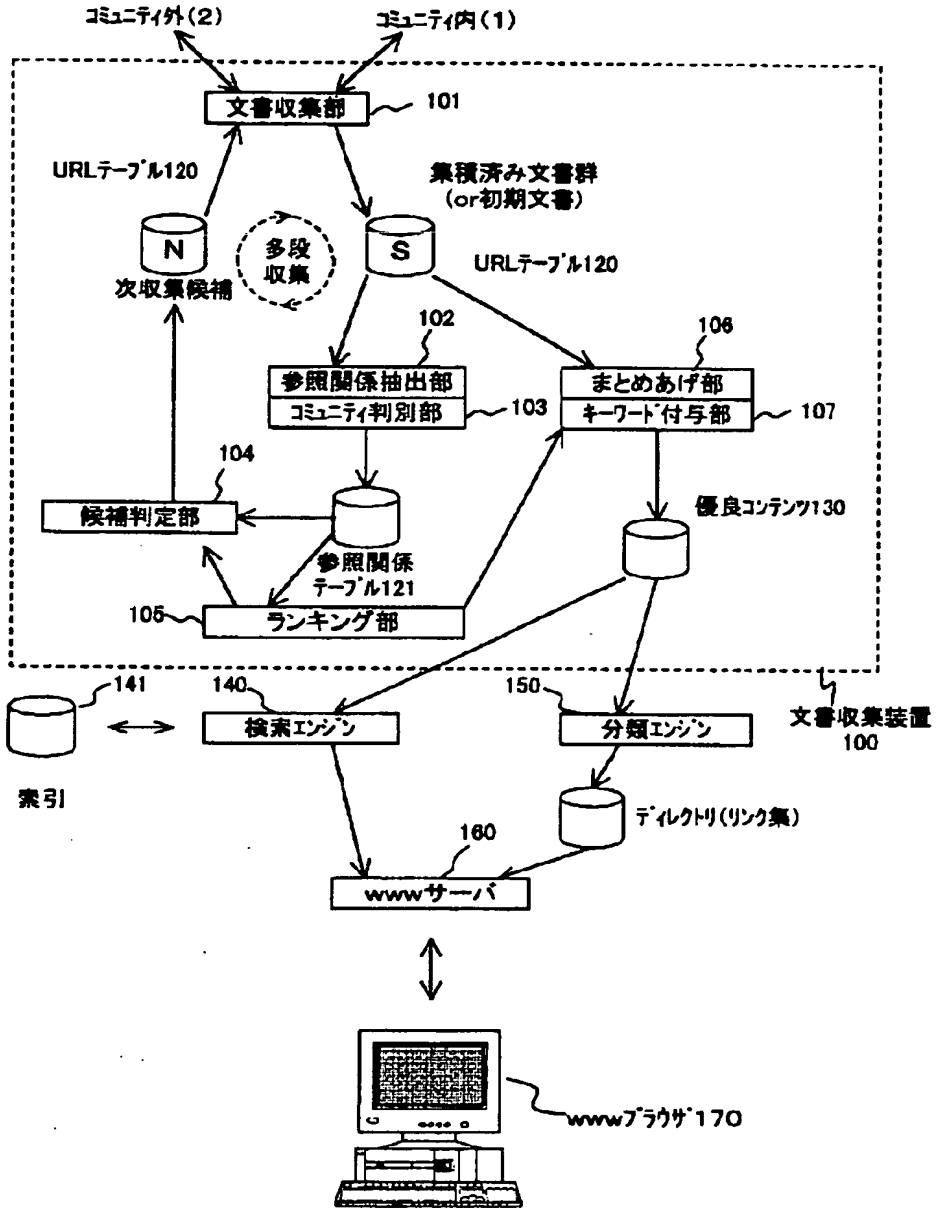
【図１】

本発明の原理図



【図２】

第1実施形態に係る文書収集装置の構成図



【図３】

URLテーブルのデータ構造の一例を示す図

URLテーブル120

文書ID	URL	集積済みフラグ	コミュニティフラグ	重要度
doc1	http://www.abc.com/	1	1	....
doc2	http://www.klm.com/	1	0	....
....	....	....	....	....

【図４】

参照関係テーブルのデータ構造の一例を示す図

参照関係テーブル121

参照元文書ID	参照先文書ID_1（コミュニティ内）	参照先文書ID_2（コミュニティ外）
doc1	doc2, doc3, doc4	doc5, doc6, doc7
doc2	doc4, doc8	doc9, doc10
....	....	....

【図５】

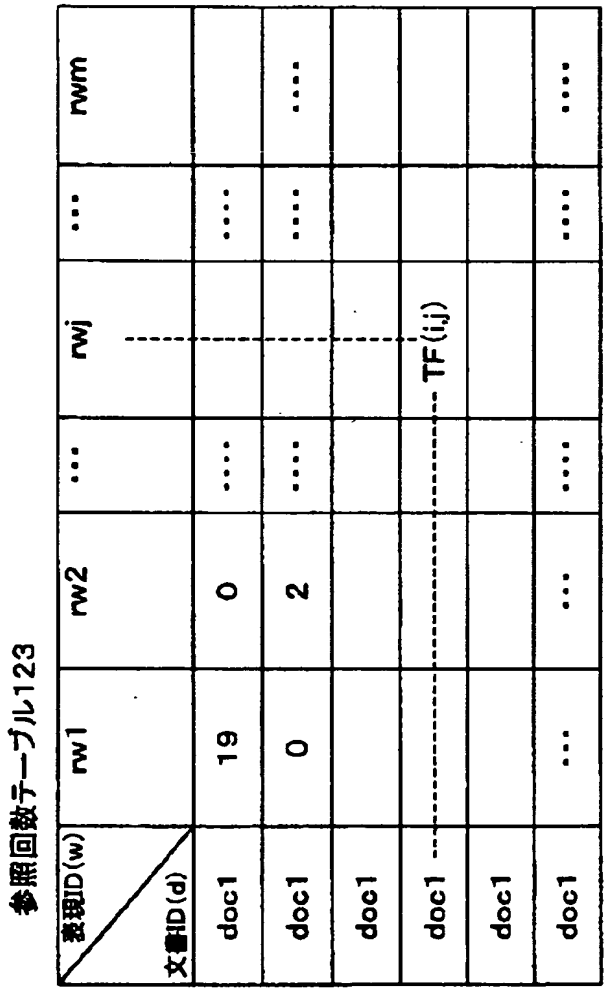
参照表現テーブルのデータ構造の一例を示す図

参照表現テーブル122

表現ID(w)	表現文字列	頻度(DF(w))	要否フラグ
rw1	ABC社	24	1
rw2	トップに戻る	28305	0
....	.....	....	....

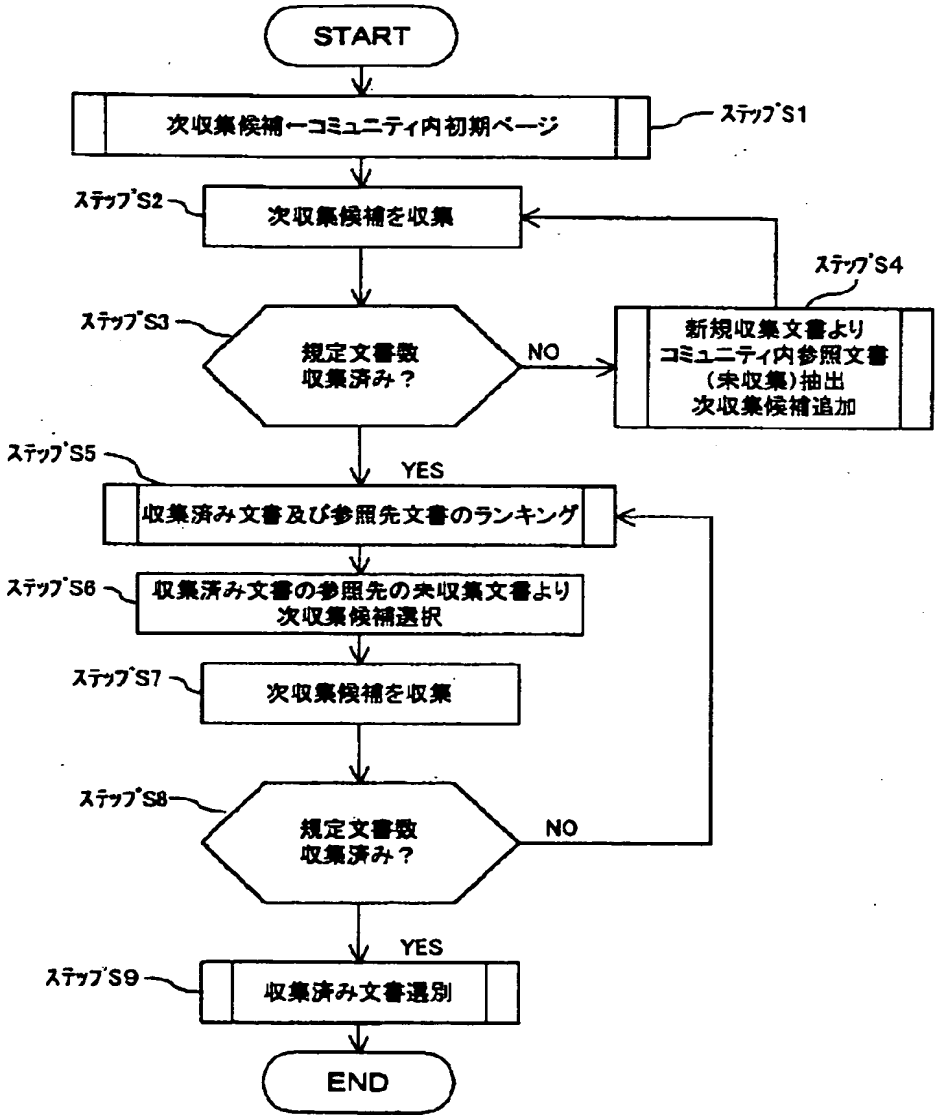
【図６】

参照回数テーブルのデータ構造の一例を示す図



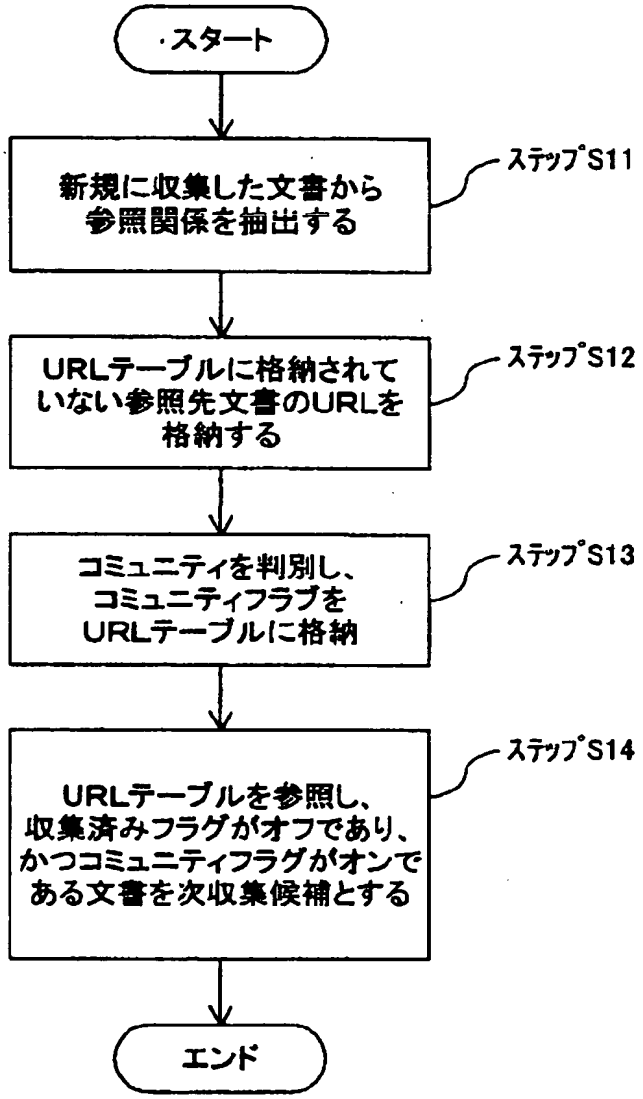
【図７】

第一実施形態に係る文書収集装置が行う処理の
大まかな流れを示すフローチャート



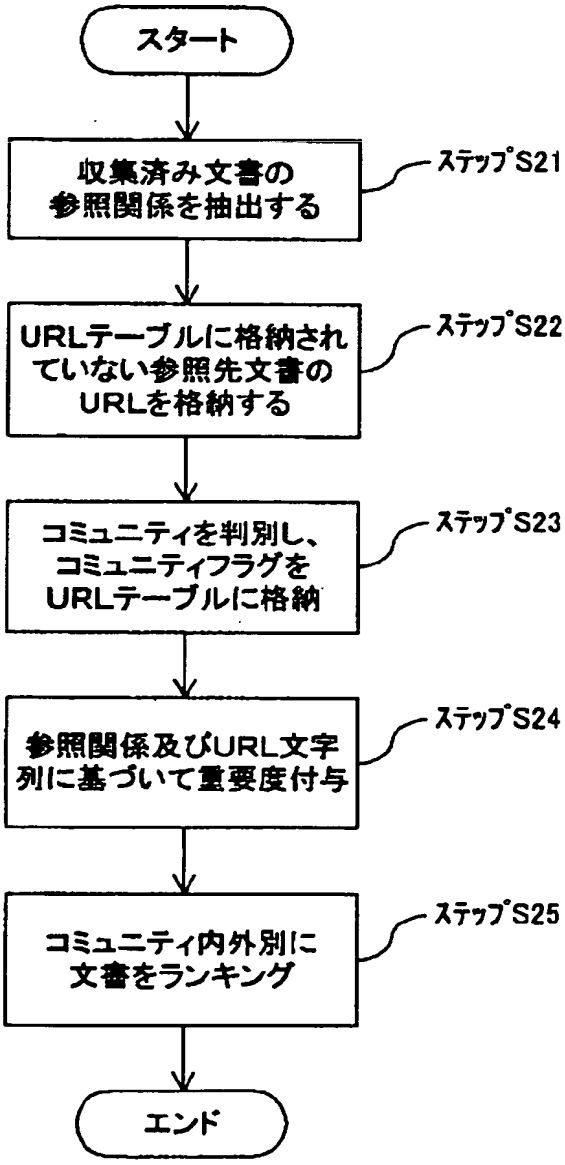
【図８】

コミュニティ内の文書を収集する際に
次収集候補を判定する処理を示すフローチャート



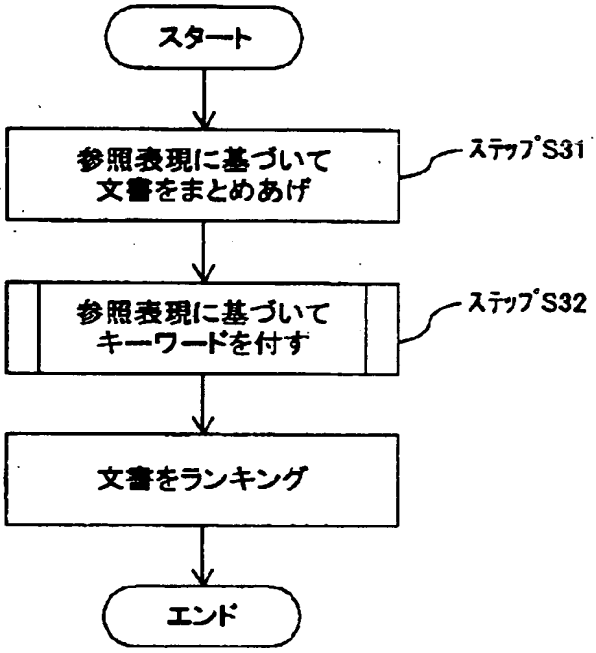
【図９】

収集済み文書及び参照先文書をランキングする処理を示すフローチャート



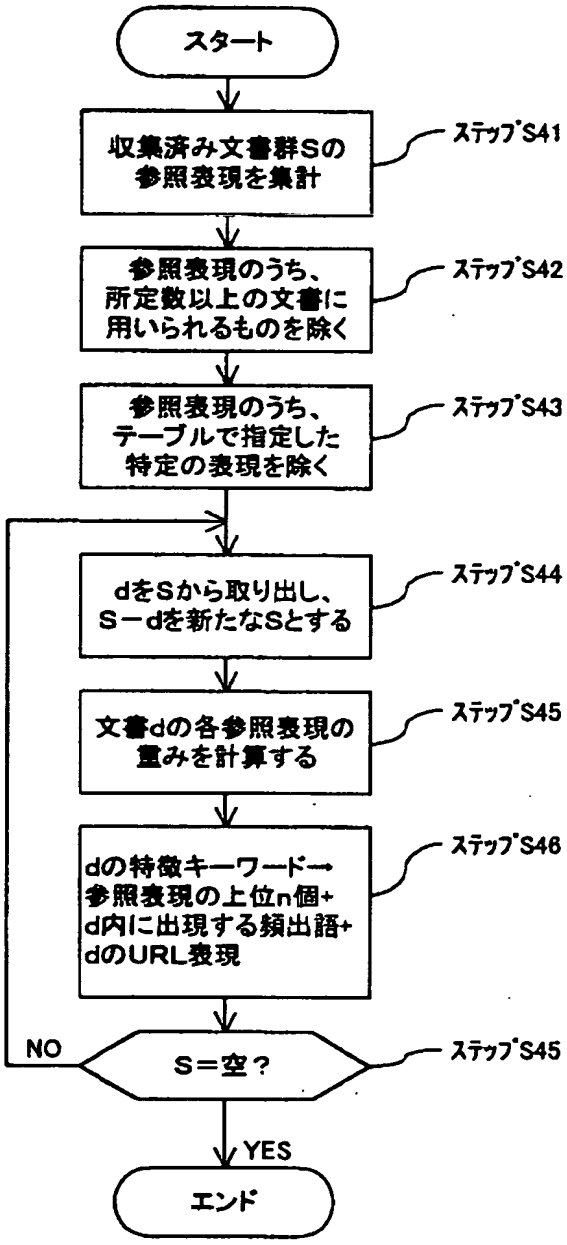
【図１０】

収集済み文書を選別する処理を示すフローチャート



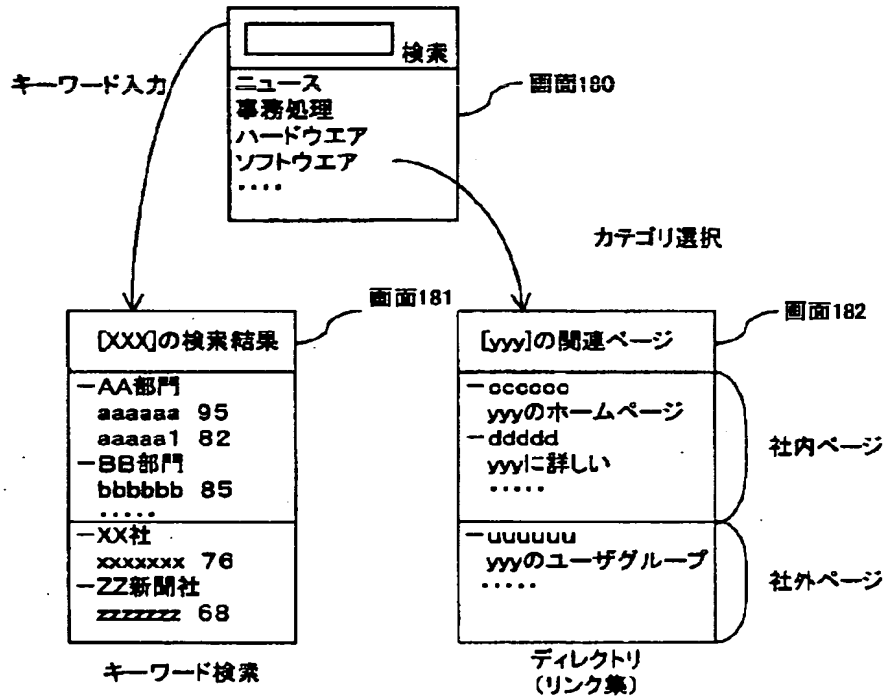
【図１１】

キーワード付与処理を示すフローチャート



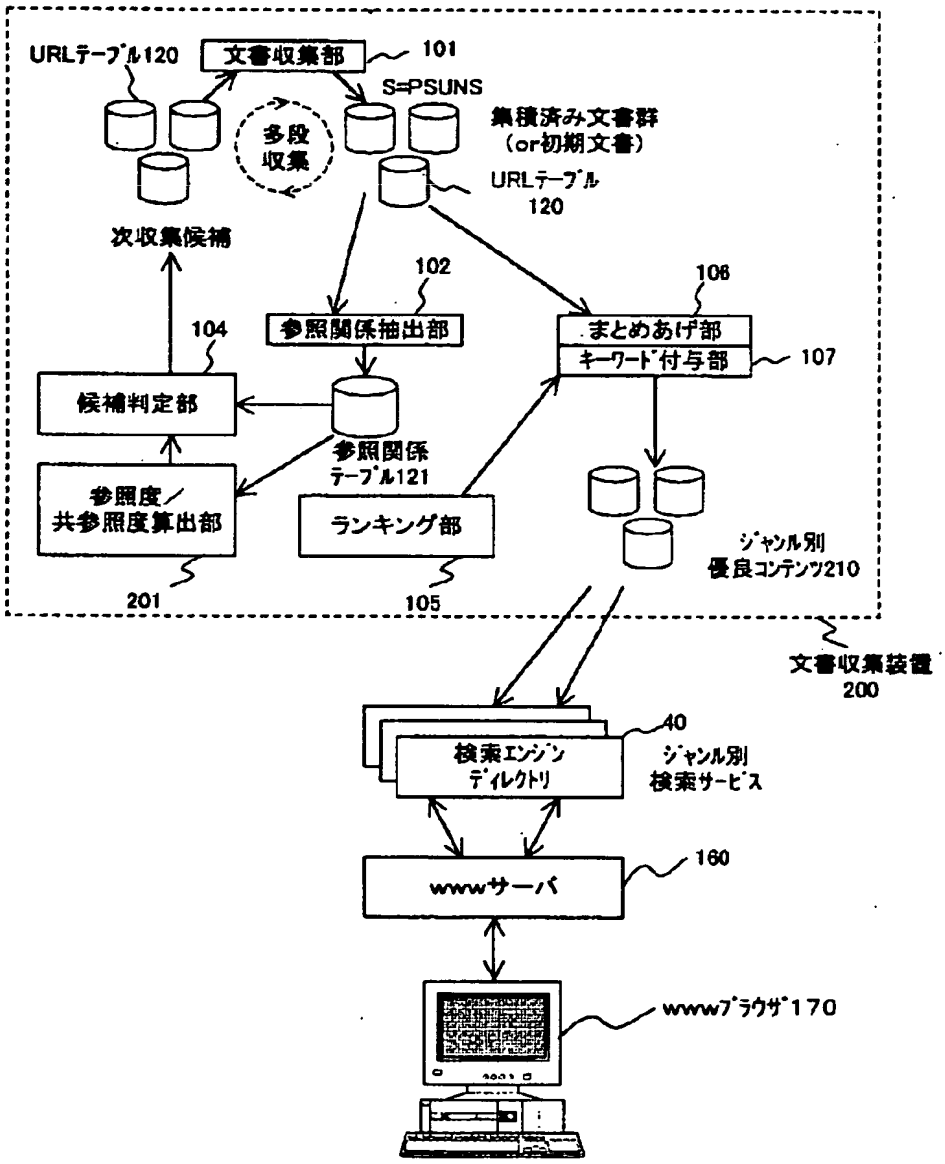
【図１２】

収集した文書を提供する画面の一例を示す図



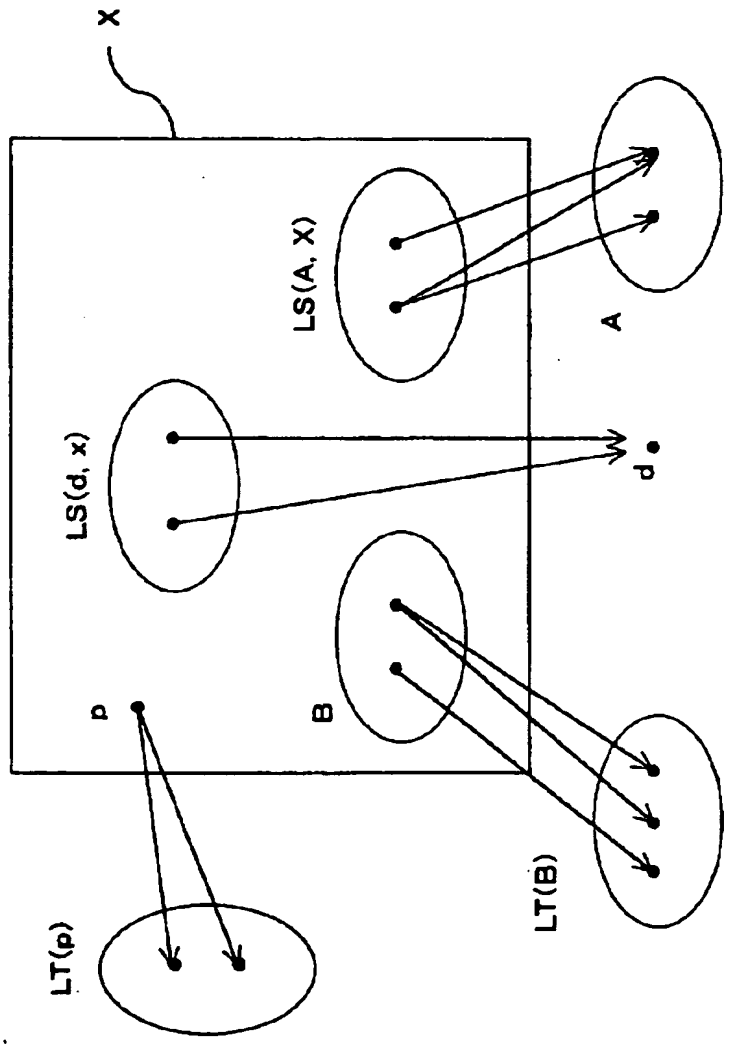
【図１３】

第2実施形態に係る文書収集装置の構成図



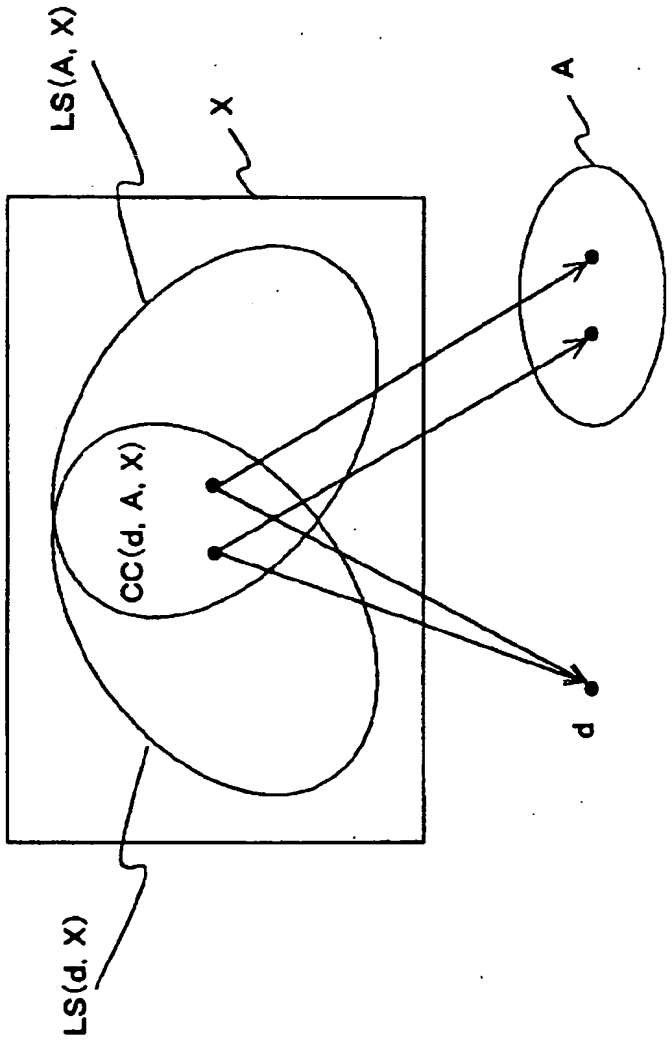
【図１４】

LT(s), LT(p), LS(d, x)及びLS(A, X)が意味する、文書の参照関係を示す図



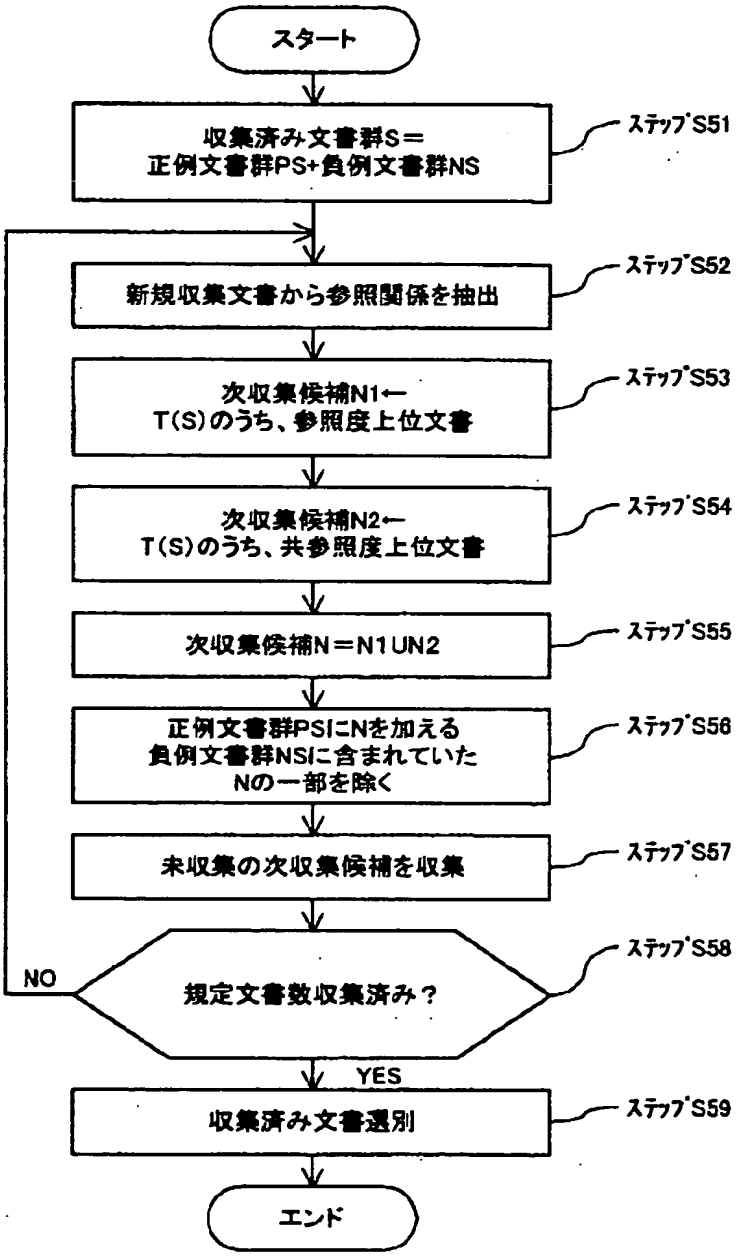
【図１５】

CC(d, A, X)が意味する文書の参照関係を示す図



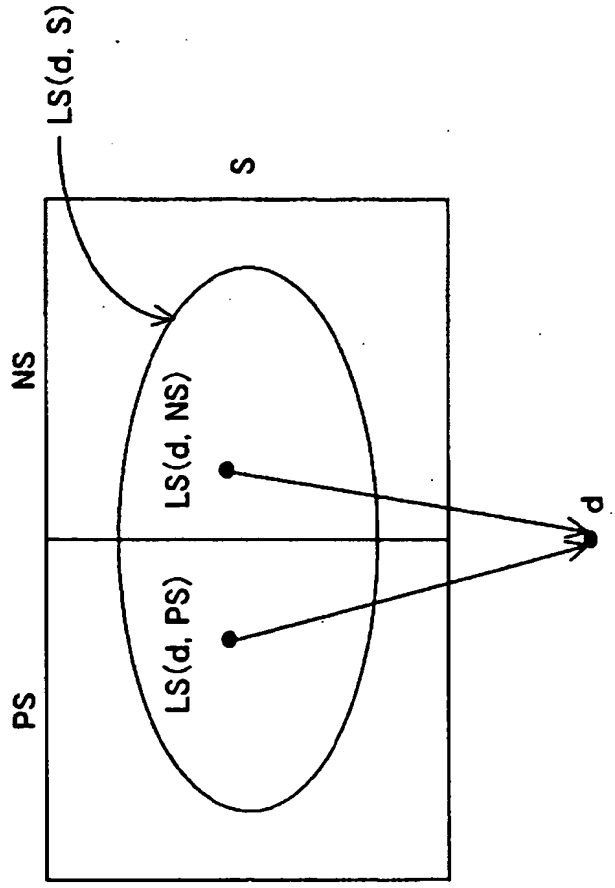
【図１６】

第2実施形態に係る文書収集装置が行う処理を示すフローチャート



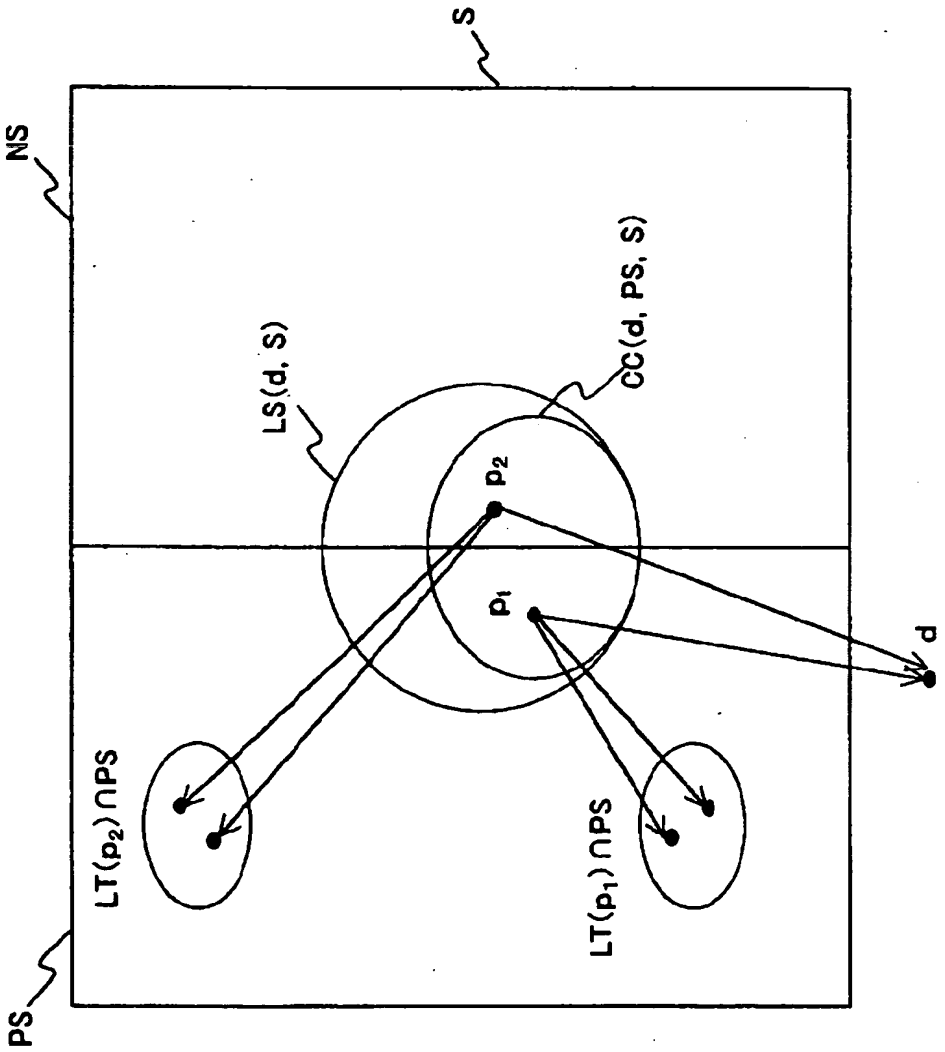
【図１７】

参照度を算出する式に含まれる各集合が意味する参照関係を示す図



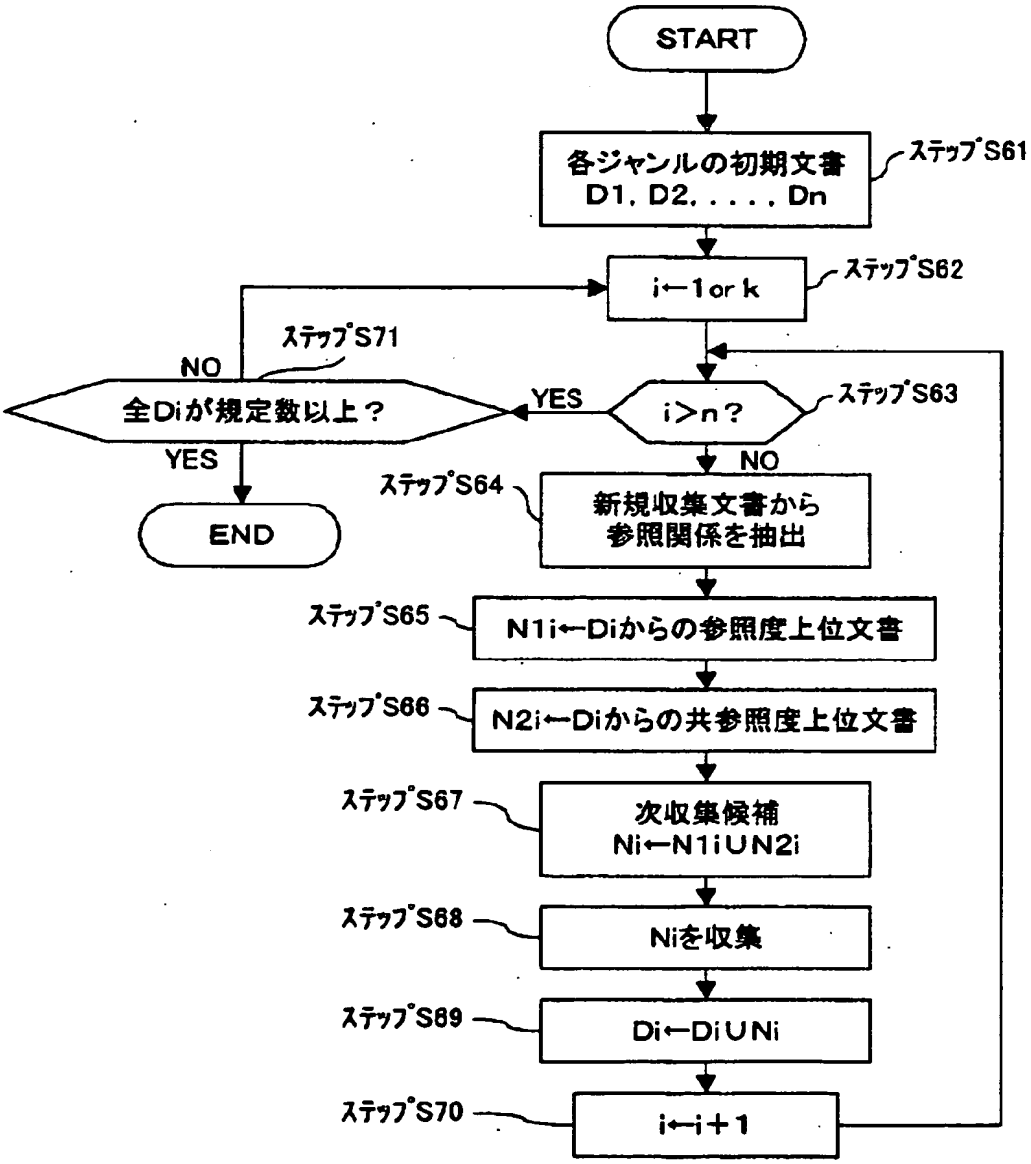
【図１８】

**共参照度を算出する式に含まれる
各集合が意味する参照関係を示す図**



【図１９】

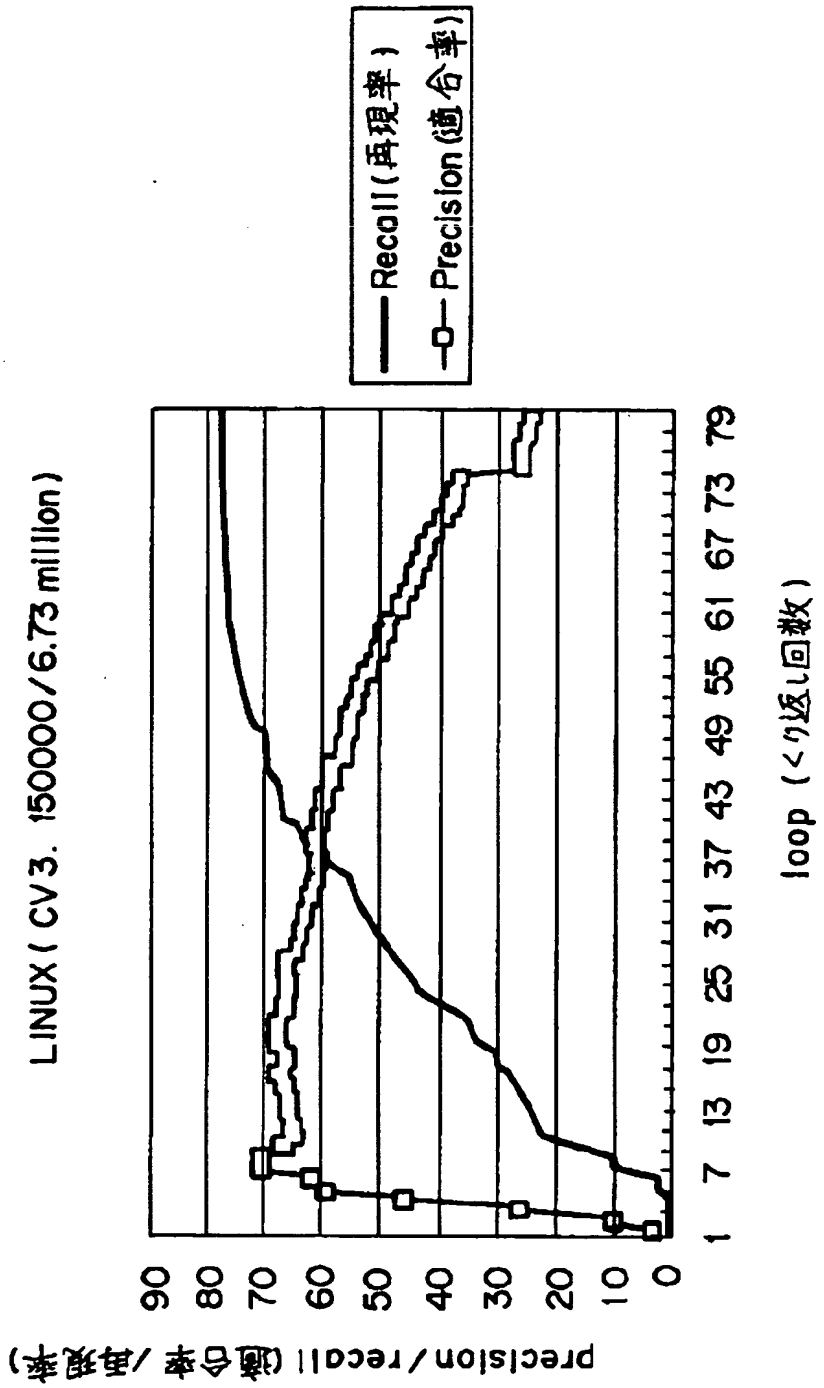
第2実施形態の変形例に係る
文書収集装置が行う処理を示すフローチャート



【図２０】

文書収集装置の収集精度の実験結果を示す図

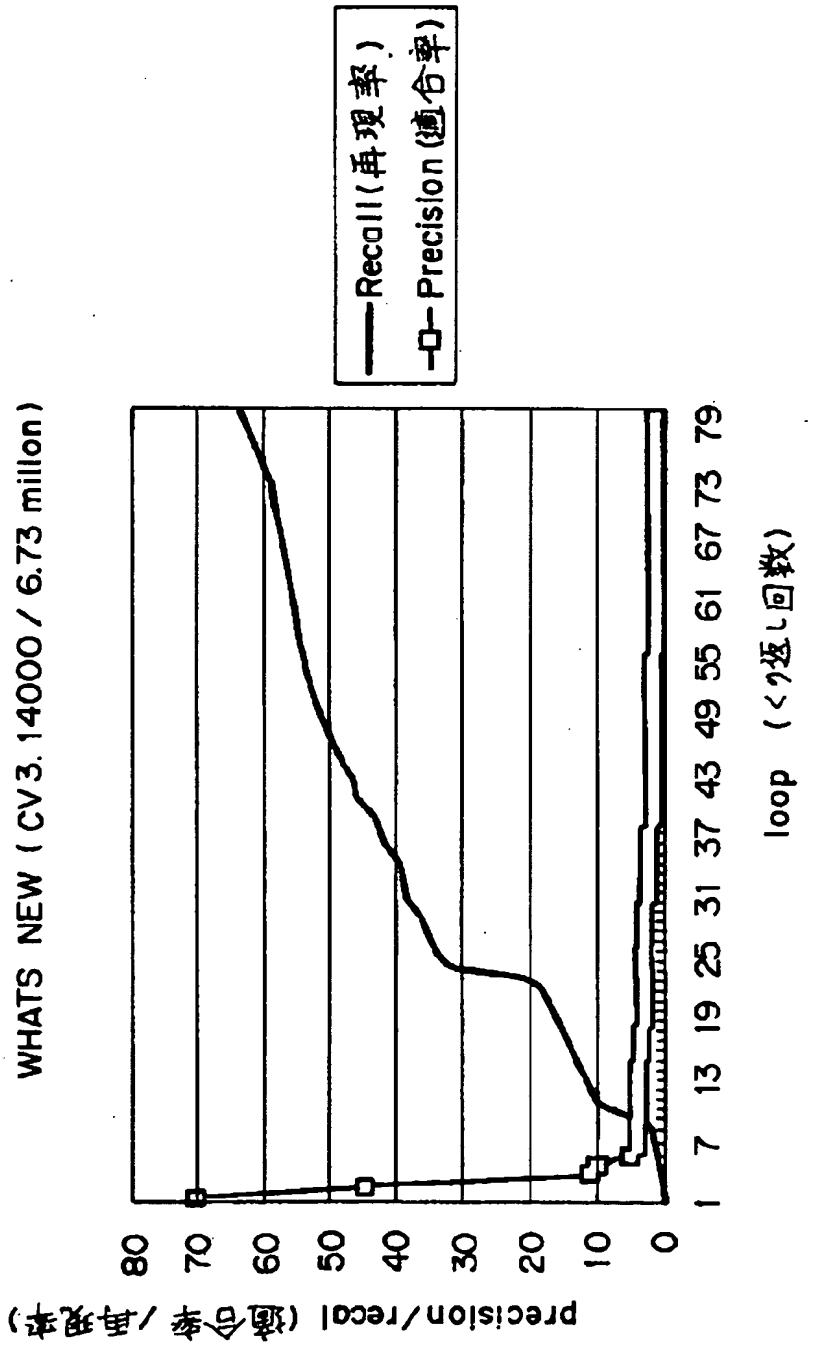
（その１）



【図２１】

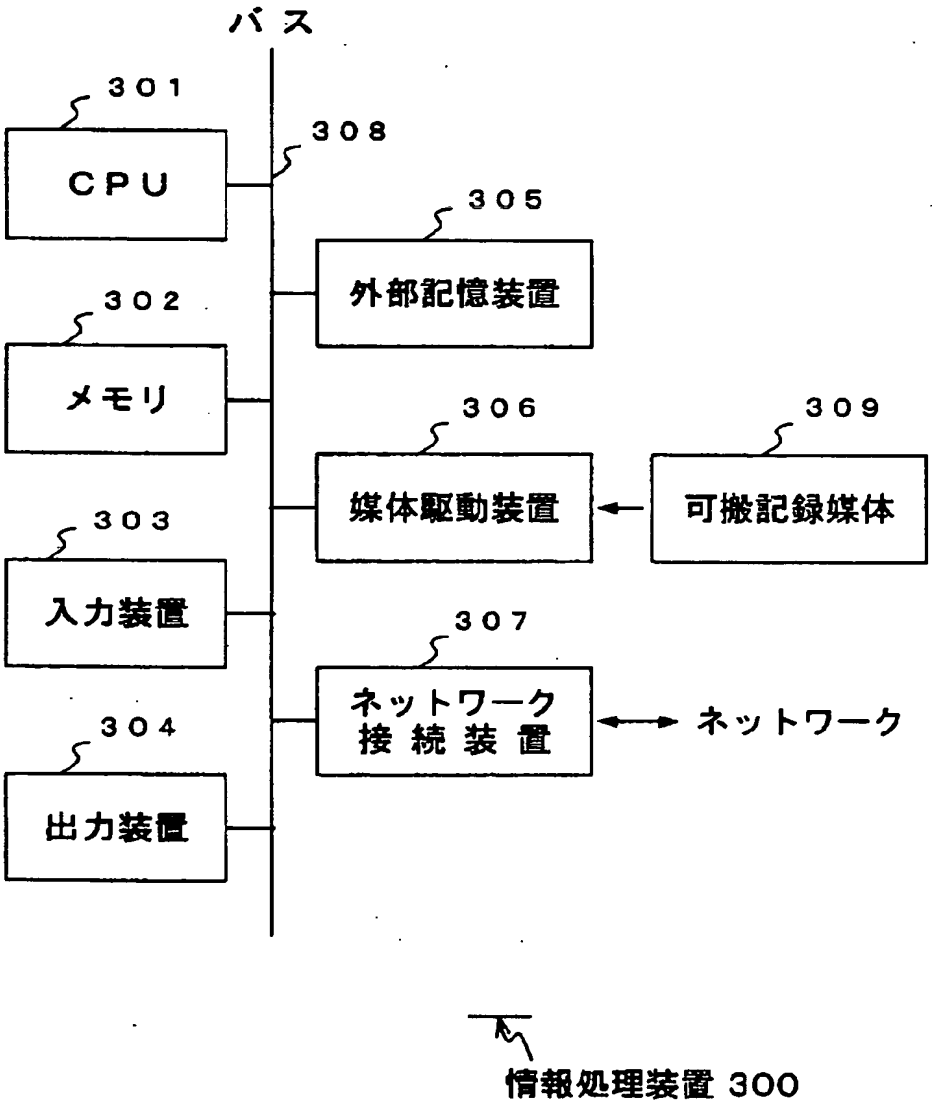
文書収集装置の収集精度の実験結果を示す図

（その２）



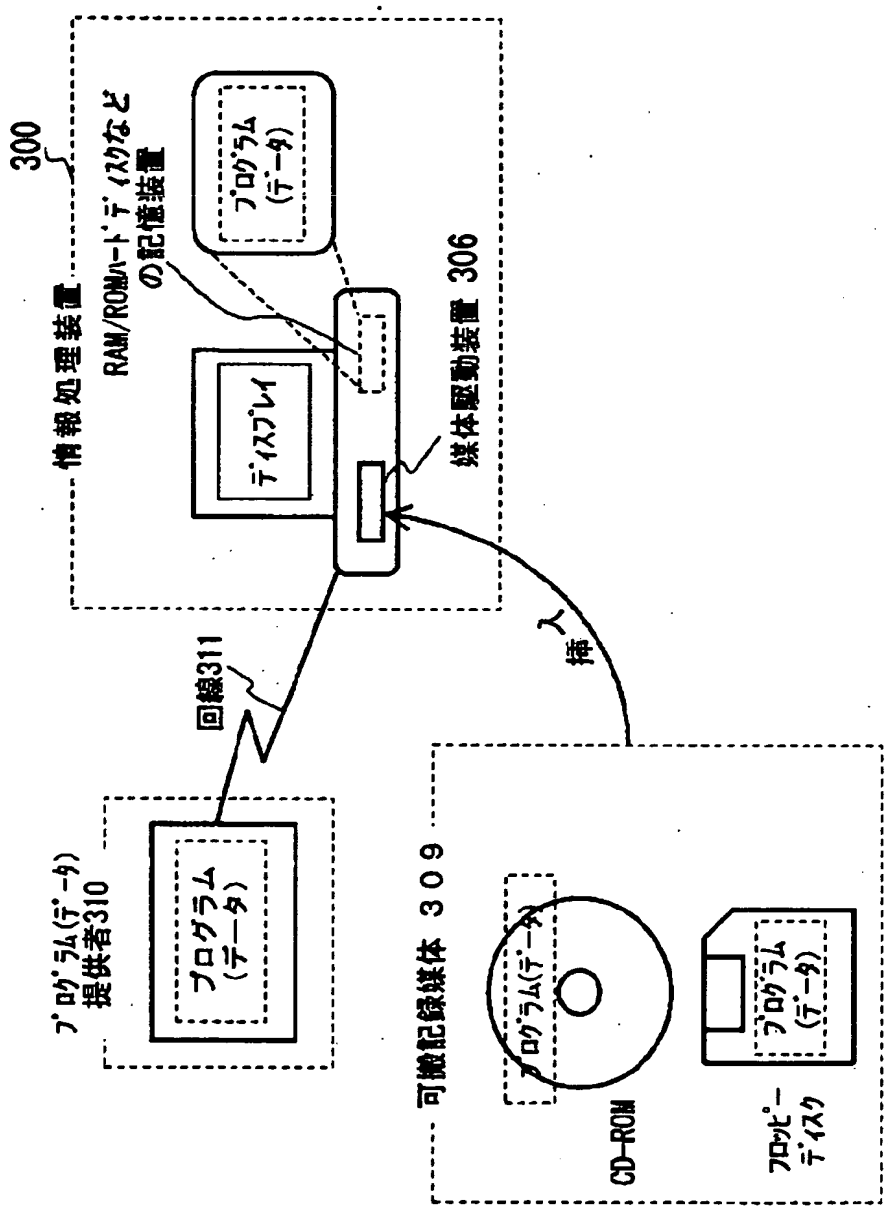
【図２２】

情　報　処　理　装　置　の　構　成　図



【図２３】

情報処理装置にプログラムやデータを供給する記憶媒体、伝送信号、及び伝送媒体を説明する図



【書類名】　　要約書

【要約】

【課題】　ネットワークから特定用途向けの文書を迅速に収集することを可能とする。

【解決手段】　ネットワークからコミュニティ向けの文書を収集する文書収集装置において、収集開始に先立って、収集の開始点となる初期文書群２０を与える。参照関係抽出手段３は、初期文書群２０から参照関係を抽出する。次候補判定手段５は、収集済みの文書群２０の参照関係に基づいて、一定の条件を満たす未収集文書を次に収集すべき文書の候補である次収集候補２１を判定する。文書収集手段２は、次収集候補２１として判定された文書を収集し、収集済み文書２０に加える。収集済み文書群２０の文書数が規定された数以上でなければ、参照関係抽出手段３は、新規に収集された文書から更に参照関係を抽出し、上述の処理を繰り返す。このように、収集済み文書群２０の文書数が規定された数以上になるまで文書の収集を繰り返す。

上記装置を、ネットワーク上のコミュニティにとって有用度の高い文書を収集するコミュニティ向けの文書収集装置として構成する場合、先に、文書収集手段２がネットワーク上のコミュニティ内から文書をまんべんなく収集した後、次候補判定手段５は、収集済み文書群２０の参照関係に基づいてコミュニティ内外の文書から次収集候補２１を決定する。

【選択図】　　図１

職権訂正履歴（職権訂正）

特許出願の番号	特願２０００－３９７９６６
受付番号	５００００１６９２２４５
書類名	特許願
担当官	濱谷　よし子　　１６１４
作成日	平成１３年　１月　４日

＜訂正内容１＞
訂正ドキュメント
明細書
訂正原因
職権による訂正
訂正メモ
【図面の簡単な説明】【図１５】を改行しました。
訂正前内容
【図１４】
ＬＴ（Ｓ）、ＬＴ（ｐ）、ＬＳ（ｄ，Ｘ）、ＬＳ（Ａ，Ｘ）が意味する文書の参照関係を示す図である。　【図１５】
ＣＣ（ｄ，Ａ，Ｘ）が意味する文書の参照関係を示す図である。
【図１６】
訂正後内容
【図１４】
ＬＴ（Ｓ）、ＬＴ（ｐ）、ＬＳ（ｄ，Ｘ）、ＬＳ（Ａ，Ｘ）が意味する文書の参照関係を示す図である。
【図１５】
ＣＣ（ｄ，Ａ，Ｘ）が意味する文書の参照関係を示す図である。
【図１６】

次頁無

出証特２００１－３０１７５６６

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　　富士通株式会社